A course in Time Series Analysis

Suhasini Subba Rao

Email: suhasini.subbarao@stat.tamu.edu

August 29, 2022

Contents

Preface

- The material for these notes come from several different places, in particular:
  - Brockwell and Davis (1998) (yellow book)
  - Shumway and Stoffer (2006) (a shortened version is Shumway and Stoffer EZ).
  - Fuller (1995)
  - Pourahmadi (2001)
  - Priestley (1983)
  - Box and Jenkins (1970)
  - Brockwell and Davis (2002) (the red book), is a very nice introduction to Time Series, which may be useful for students who don't have a rigourous background in mathematics.
  - Wilson Tunnicliffe et al. (2020)
  - Tucker and Politis (2021)
  - A whole bunch of articles.
  - My own random thoughts and derivations.
- Tata Subba Rao and Piotr Fryzlewicz were very generous in giving advice and sharing homework problems.
- When doing the homework, you are encouraged to use all materials available, including Wikipedia, Mathematica/Maple (software which allows you to easily derive analytic expressions, a web-based version which is not sensitive to syntax is Wolfram-alpha).

- You are encouraged to use `R` (see David Stoffer's tutorial). I have tried to include Rcode in the notes so that you can replicate some of the results.

- Exercise questions will be in the notes and will be set at regular intervals.

- Finally, these notes are dedicated to my wonderful Father, whose inquisitive questions, and unconditional support inspired my quest in time series.

Chapter 1

Introduction

A time series is a series of observations x_t, observed over a period of time. Typically the observations can be over an entire interval, randomly sampled on an interval or at fixed time points. Different types of time sampling require different approaches to the data analysis.

In this course we will focus on the case that observations are observed at fixed equidistant time points, hence we will suppose we observe $\{x_t : t \in \mathbb{Z}\}$ ($\mathbb{Z} = \{\ldots, 0, 1, 2 \ldots\}$).

Let us start with a simple example, independent, uncorrelated random variables (the simplest example of a time series). A plot is given in Figure 1.1. We observe that there aren't any clear patterns in the data. Our best forecast (predictor) of the next observation is zero (which appears to be the mean). The feature that distinguishes a time series from classical statistics is that there is dependence in the observations. This allows us to obtain better forecasts of future observations. Keep Figure 1.1 in mind, and compare this to the following real examples of time series (observe in all these examples you see patterns).

1.1 Time Series data

Below we discuss four different data sets.

The Southern Oscillation Index from 1876-present

The Southern Oscillation Index (SOI) is an indicator of intensity of the El Nino effect (see wiki). The SOI measures the fluctuations in air surface pressures between Tahiti and Darwin.

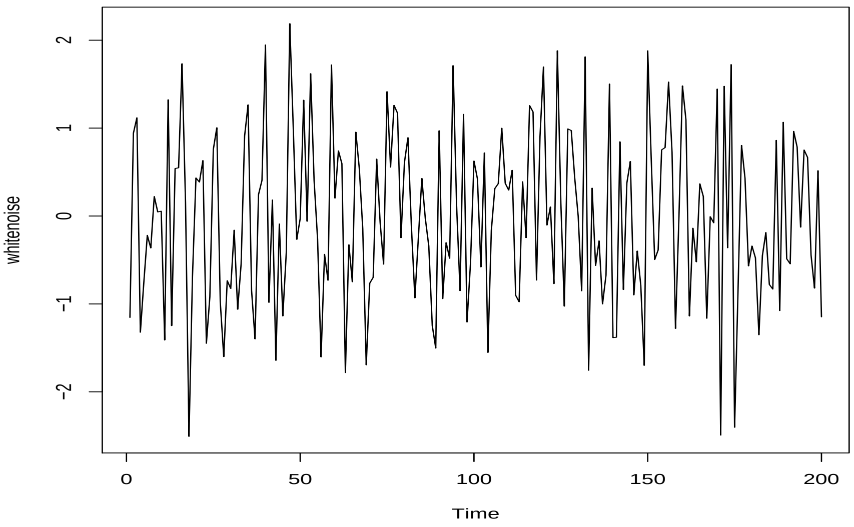


Figure 1.1: Plot of independent uncorrelated random variables

In Figure 1.2 we give a plot of monthly SOI from January 1876 - July 2014 (note that there is some doubt on the reliability of the data before 1930). The data was obtained from `http://www.bom.gov.au/climate/current/soihtm1.shtml`. Using this data set one major goal is to look for patterns, in particular periodicities in the data.

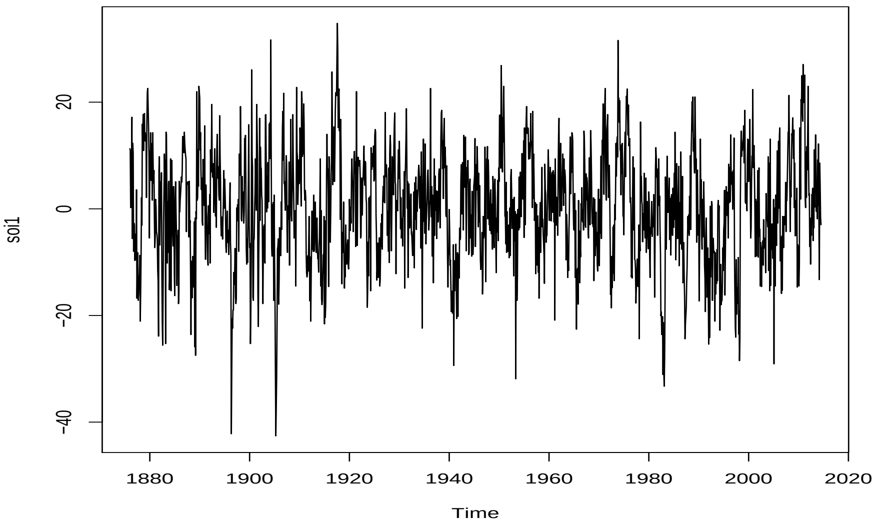


Figure 1.2: Plot of monthly Southern Oscillation Index, 1876-2014

Nasdaq Data from 1985-present

The daily closing Nasdaq price from 1st October, 1985- 8th August, 2014 is given in Figure 1.3. The (historical) data was obtained from `https://uk.finance.yahoo.com`. See also `http://www.federalreserve.gov/releases/h10/Hist/`. Of course with this type of data the goal is to make money! Therefore the main object is to forecast (predict future volatility).

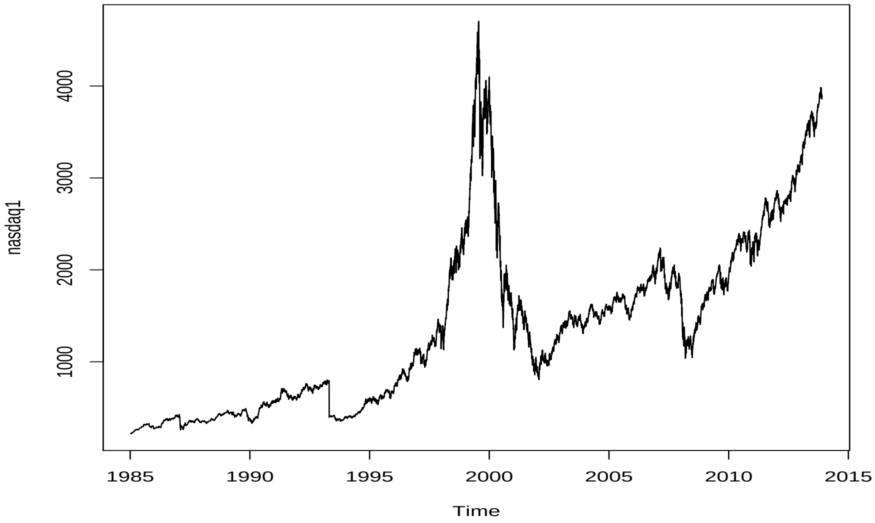


Figure 1.3: Plot of daily closing price of Nasdaq 1985-2014

Yearly sunspot data from 1700-2013

Sunspot activity is measured by the number of sunspots seen on the sun. In recent years it has had renewed interest because times in which there are high activity causes huge disruptions to communication networks (see wiki and NASA).

In Figure 1.4 we give a plot of yearly sunspot numbers from 1700-2013. The data was obtained from `http://www.sidc.be/silso/datafiles`. For this type of data the main aim is to both look for patterns in the data and also to forecast (predict future sunspot activity).

Yearly and monthly average temperature data

Given that climate change is a very topical subject we consider global temperature data. Figure 1.5 gives the yearly temperature anomalies from 1880-2013 and in Figure 1.6 we plot

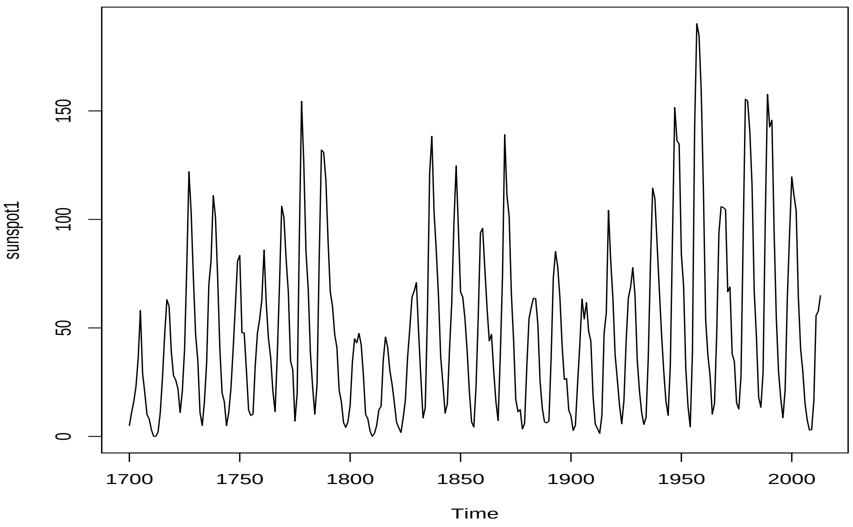


Figure 1.4: Plot of Sunspot numbers 1700-2013

the monthly temperatures from January 1996 - July 2014. The data was obtained from `http://data.giss.nasa.gov/gistemp/graphs_v3/Fig.A2.txt` and `http://data.giss.nasa.gov/gistemp/graphs_v3/Fig.C.txt` respectively. For this type of data one may be trying to detect for global warming (a long term change/increase in the average temperatures). This would be done by fitting trend functions through the data. However, sophisticated time series analysis is required to determine whether these estimators are statistically significant.

1.2 `R` code

A large number of the methods and concepts will be illustrated in `R`. If you are not familar with this language please learn the basics.

Here we give the `R` code for making the plots above.

```
# assuming the data is stored in your main directory we scan the data into R
soi <-  scan("~/soi.txt")
soi1 <-  ts(monthlytemp,start=c(1876,1),frequency=12)
# the function ts creates a timeseries object, start = starting year,
```

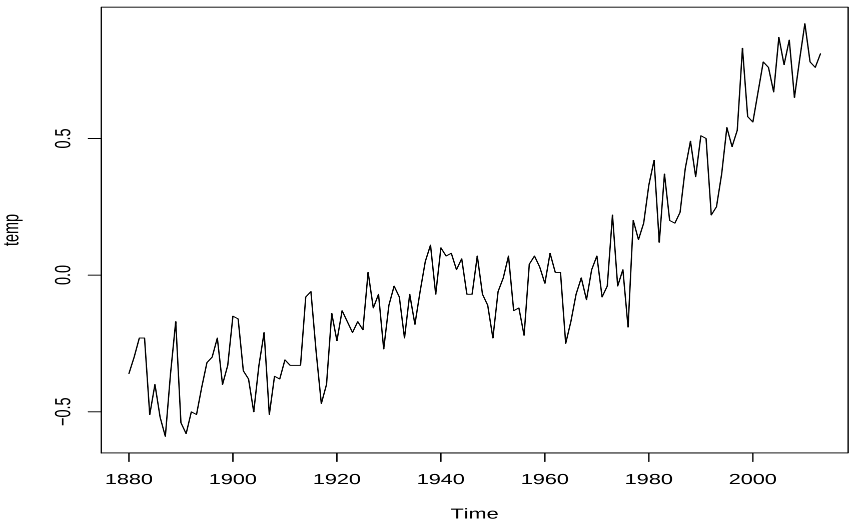


Figure 1.5: Plot of global, yearly average, temperature anomalies, 1880 - 2013

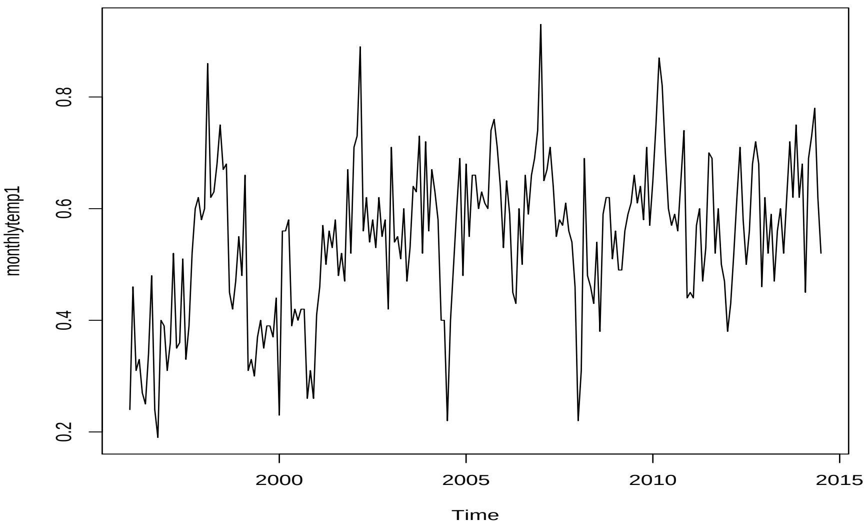


Figure 1.6: Plot of global, monthly average, temperatures January, 1996 - July, 2014.

```
# where 1 denotes January. Frequency = number of observations in a
# unit of time (year). As the data is monthly it is 12.
plot.ts(soi1)
```

Dating plots properly is very useful. This can be done using the package `zoo` and the function `as.Date`.

1.3 Filtering time series

Often we transform data to highlight features or remove unwanted features. This is often done by taking the log transform or a linear transform.

It is no different for time series. Often a transformed time series can be easier to analyse or contain features not apparent in the original time series. In these notes we mainly focus on *linear* transformation of the time series. Let $\{X_t\}$ denote the original time series and $\{Y_t\}$ transformed time series where

$$Y_t = \sum_{j=-\infty}^{\infty} h_j X_{t-j}$$

where $\{h_j\}$ are weights.

In these notes we focus on two important types of linear transforms of the time series:

(i) Linear transforms that can be used to estimate the underlying mean function.

(ii) Linear transforms that allow us to obtain a deeper understanding on the actual stochastic/random part of the observed time series.

In the next chapter we consider estimation of a time-varying mean in a time series and will use some of the transforms alluded to above.

1.4 Terminology

- iid (independent, identically distributed) random variables. The simplest time series you could ever deal with!

Chapter 2

Trends in a time series

Objectives:

- Parameter estimation in parametric trend.
- The Discrete Fourier transform.
- Period estimation.

In time series, the main focus is on understanding and modelling the relationship between observations. A typical time series model looks like

$$Y_t = \mu_t + \varepsilon_t,$$

where μ_t is the underlying mean and ε_t are the residuals (errors) which the mean cannot explain. Formally, we say $\mathrm{E}[Y_t] = \mu_t$. We will show later in this section, that when data it can be difficult to disentangle to the two. However, a time series analysist usually has a few jobs to do when given such a data set. Either (a) estimate μ_t, we discuss various methods below, this we call $\widehat{\mu}_t$ or (b) transform $\{Y_t\}$ in such a way that μ_t "disappears". What method is used depends on what the aims are of the analysis. In many cases it is to estimate the mean μ_t. But the *e*stimated residuals

$$\widehat{\varepsilon}_t = Y_t - \widehat{\mu}_t$$

also plays an important role. By modelling $\{\varepsilon_t\}_t$ we can understand its dependence structure. This knowledge will allow us to construct reliable confidence intervals for the mean μ_t. Thus the residuals $\{\varepsilon_t\}_t$ play an important but peripheral role. However, for many data sets the residuals $\{\varepsilon_t\}_t$ are important and it is the mean that is a nuisance parameters. In such situations we either find a transformation which removes the mean and focus our analysis on the residuals ε_t. The main focus of this class will be on understanding the structure of the residuals $\{\varepsilon_t\}_t$. However, in this chapter we study ways in which to estimate the mean μ_t.

Shumway and Stoffer, Chapter 2, and Brockwell and Davis (2002), Chapter 1.

2.1 Parametric trend

In many situations, when we observe time series, regressors are also available. The regressors may be an exogenous variable but it could even be time (or functions of time), since for a time series the index t has a meaningful ordering and can be treated as a regressor. Often the data is assumed to be generated using a parametric model. By parametric model, we mean a model where all but a finite number of parameters is assumed known. Possibly, the simplest model is the linear model. In time series, a commonly used linear model is

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t, \tag{2.1}$$

or

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t, \tag{2.2}$$

where β_0, β_1 and β_2 are unknown. These models are *l*inear because they are linear in the regressors. An example of a popular nonlinear models is

$$Y_t = \frac{1}{1 + \exp[\beta_0(t - \beta_1)]} + \varepsilon_t. \tag{2.3}$$

where β_0 and β_1 are unknown. As the parameters in this model are *inside* a function, this

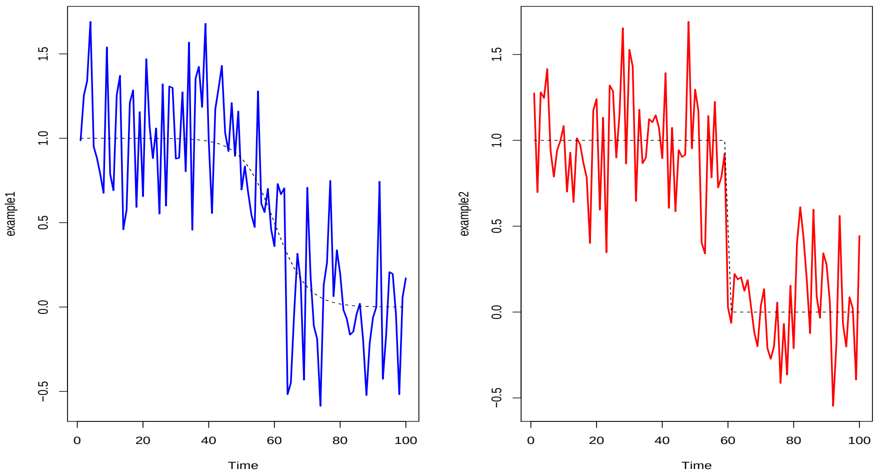


Figure 2.1: The function Y_t in (2.3) with iid noise with $\sigma = 0.3$. Dashed is the truth. Left: $\beta_0 = 0.2$ and $\beta_1 = 60$. Right: $\beta_0 = 5$ and $\beta_1 = 60$

is an example of a nonlinear model. The above nonlinear model (called a smooth transition model), is used to model transitions from one state to another (as it is monotonic, increasing or decreasing depending on the sign of β_0). Another popular model for modelling ECG data is the burst signal model (see Swagata Nandi et. al.)

$$Y_t = A \exp\left(\beta_0(1 - \cos(\beta_2 t))\right) \cdot \cos(\theta t) + \varepsilon_t \tag{2.4}$$

Both these nonlinear parametric models motivate the general nonlinear model

$$Y_t = g(\underline{x}_t, \theta) + \varepsilon_t, \tag{2.5}$$

where $g(\underline{x}_t, \theta)$ is the nonlinear trend, g is a known function but θ is unknown. Observe that most models include an additive noise term $\{\varepsilon_t\}_t$ to account for variation in Y_t that the trend cannot explain.

Real data example Monthly temperature data. This time series appears to include seasonal behaviour (for example the southern oscillation index). Seasonal behaviour is often modelled

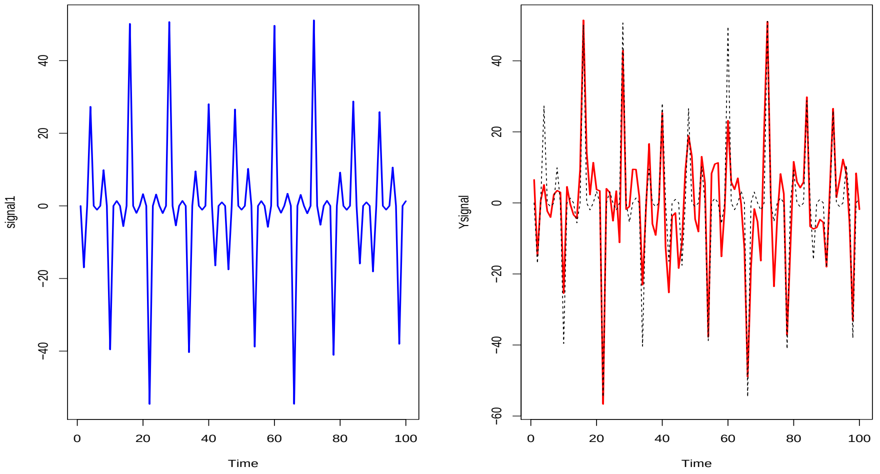


Figure 2.2: The Burst signal (equation (2.4)) $A = 1$, $\beta_0 = 2$, $\beta_1 = 1$ and $\theta = \pi/2$ with iid noise with $\sigma = 8$. Dashed is the truth. Left: True Signal. Right: True Signal with noise

with sines and cosines

$$Y_t = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{P}\right) + \beta_3 \cos\left(\frac{2\pi t}{P}\right) + \varepsilon_t,$$

where P denotes the length of the period. If P is known, for example there are 12 months in a year so setting $P = 12$ is sensible. Then we are modelling trends which repeat every 12 months (for example monthly data) and

$$Y_t = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{12}\right) + \beta_3 \cos\left(\frac{2\pi t}{12}\right) + \varepsilon_t. \qquad (2.6)$$

is an example of a $l$inear model.

On the other hand, if P is known and has to be estimated from the data too. Then this is an example of a $n$onlinear model. We consider more general periodic functions in Section 2.5.

2.1.1 Least squares estimation

In this section we review simple estimation methods. In this section, we do not study the properties of these estimators. We touch on that in the next chapter.

A quick review of least squares Suppose that variable X_i are believed to influence the response variable Y_i. So far the relationship is unknown, but we regress (project $\underline{Y}_n = (Y_1, \ldots, Y_n)')$ onto $\underline{X}_n = (X_1, \ldots, X_n)$ using least squares. We know that this means finding the α which minimises the distance

$$\sum_{i=1}^{n}(Y_i - \alpha X_i)^2.$$

The α, which minimises the above, for mathematical convenience we denote as

$$\widehat{\alpha}_n = \arg\min_{\alpha} \sum_{i=1}^{n}(Y_i - \alpha X_i)^2$$

and it has an analytic solution

$$\widehat{\alpha}_n = \frac{\langle \underline{Y}_n, \underline{X}_n \rangle}{\|\underline{X}_n\|_2^2} = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}.$$

A geometric interpretation is that the vector $\underline{Y}_n$ is projected onto $\underline{X}_n$ such that

$$\underline{Y}_n = \widehat{\alpha}_n \underline{X}_n + \underline{\varepsilon}_n$$

where $\underline{\varepsilon}_n$ is orthogonal to $\underline{X}_n$ in other words

$$\langle \underline{X}_n, \underline{\varepsilon}_n \rangle = \sum_{i=1}^{n} X_i \varepsilon_{i,n} = 0.$$

But so far no statistics. We can always project a vector on another vector. We have made no underlying assumption on what generates Y_i and how X_i really impacts X_i. Once we do this we are in the realm of modelling. We do this now. Let us suppose the **data generating process** (often abbreviated to DGP) is

$$Y_i = \alpha X_i + \varepsilon_i,$$

here we place the orthogonality assumption between X_i and ε_i by assuming that they are uncorrelated i.e. $\text{cov}[\varepsilon_i, X_i]$. This basically means ε_i contains no linear information about X_i. Once a model has been established. We can make more informative statements about $\widehat{\alpha}_n$. In this case $\widehat{\alpha}_n$ is estimating α and $\widehat{\alpha}_n X_i$ is an estimator of the mean αX_i.

Multiple linear regression The above is regress $\underline{Y}_n$ onto just one regressor $\underline{X}_n$. Now consider regressing $\underline{Y}_n$ onto several regressors $(\underline{X}_{1,n}, \ldots, \underline{X}_{p,n})$ where $\underline{X}'_{i,n} = (X_{i,1}, \ldots, X_{i,n})$. This means projecting $\underline{Y}_n$ onto several regressors $(\underline{X}_{1,n}, \ldots, \underline{X}_{p,n})$. The coefficients in this projection are $\widehat{\underline{\alpha}}_n$, where

$$\begin{aligned}
\widehat{\underline{\alpha}}_n &= \arg\min_{\alpha} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \alpha_j X_{i,j})^2 \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y}_n.
\end{aligned}$$

and $\mathbf{X} = (\underline{X}_{1,n}, \ldots, \underline{X}_{p,n})$. If the vectors $\{\underline{X}_{j,n}\}_{j=1}^{p}$ are orthogonal, then $\mathbf{X}'\mathbf{X}$ is diagonal matrix. Then the expression for $\widehat{\underline{\alpha}}_n$ can be simplified

$$\widehat{\alpha}_{j,n} = \frac{\langle \underline{Y}_n, \underline{X}_{j,n} \rangle}{\|\underline{X}_{j,n}\|_2^2} = \frac{\sum_{i=1}^{n} Y_i X_{i,j}}{\sum_{i=1}^{n} X_{i,j}^2}.$$

Orthogonality of regressors is very useful, it allows simple estimation of parameters and avoids issues such as collinearity between regressors.

Of course we can regress $\underline{Y}_n$ onto anything. In order to make any statements at the population level, we have to make an assumption about the true relationship between Y_i and $\underline{X}'_{i,n} = (X_{i,1}, \ldots, X_{i,p})$. Let us suppose the data generating process is

$$Y_i = \sum_{j=1}^{p} \alpha_j X_{i,j} + \varepsilon_i.$$

Then $\widehat{\underline{\alpha}}_n$ is an estimator of $\underline{\alpha}$. But how good an estimator it is depends on the properties of $\{\varepsilon_i\}_{i=1}^{n}$. Typically, we make the assumption that $\{\varepsilon_i\}_{i=1}^{n}$ are independent, identically distributed random variables. But if Y_i is observed over time, then this assumption may well be untrue (we come to this later and the impact it may have).

If there is a choice of many different variables, the AIC (Akaike Information Criterion) is usually used to select the important variables in the model (see wiki).

Nonlinear least squares Least squares has a nice geometric interpretation in terms of projections. But for models like (2.3) and (2.4) where the unknown parameters are not the coefficients of the regressors ($Y_i = g(\underline{X}_i, \theta) + \varepsilon_i$), least squares can still be used to estimate θ

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (Y_i - g(\underline{X}_i, \theta))^2.$$

Usually, for nonlinear linear least squares no analytic solution for $\widehat{\theta}_n$ exists and one has to use a numerical routine to minimise the least squares criterion (such as `optim` in R). These methods can be highly sensitive to initial values (especially when there are many parameters in the system) and may only give the local minimum. However, in some situations one by "clever" manipulations one can find simple methods for minimising the above.

Again if the true model is $Y_i = g(\underline{X}_i, \theta) + \varepsilon_i$, then $\widehat{\theta}_n$ is an estimator of θ.

2.2 Differencing

Let us return to the Nasdaq data (see Figure 1.3). We observe what appears to be an upward trend. First differencing often removes the trend in the model. For example if $Y_t = \beta_0 + \beta_1 t + \varepsilon_t$, then

$$Z_t = Y_{t+1} - Y_t = \beta_1 + \varepsilon_{t+1} - \varepsilon_t.$$

Another model where first difference is also extremely helpful are those which have a stochastic trend. A simple example is

$$Y_t = Y_{t-1} + \varepsilon_t, \tag{2.7}$$

where $\{\varepsilon_t\}_t$ are iid random variables. It is believed that the logorithm of the Nasdaq index data (see Figure 1.3 is an example of such a model). Again by taking first differences we have

$$Z_t = Y_{t+1} - Y_t = \varepsilon_{t+1}.$$

Higher order differences Taking higher order differences can remove higher order polynomials and stochastic trends. For example if $Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$ then

$$Z_t^{(1)} = Y_{t+1} - Y_t = \beta_1 + 2\beta_2 t + \varepsilon_{t+1} - \varepsilon_t,$$

this still contains the trend. Taking second differences removes that

$$Z_t^{(2)} = Z_t^{(1)} - Z_{t-1}^{(1)} = 2\beta_2 + \varepsilon_{t+1} - 2\varepsilon_t + \varepsilon_{t-1}.$$

In general, the number of differences corresponds to the order of the polynomial. Similarly if a stochastic trend is of the form

$$Y_t = 2Y_{t-1} - Y_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}_t$ are iid. Then second differencing will return us to ε_t.

Warning Taking too many differences can induce "ugly" dependences in the data. This happens with the linear trend model $Y_t = \beta_0 + \beta_1 t + \varepsilon_t$ when we difference $\{Y_t\}$ is independent over time but $Z_t = Y_t - Y_{t-1} = \beta_1 + \varepsilon_{t+1} - \varepsilon_t$ is dependent over time since

$$Z_t = \beta_1 + \varepsilon_{t+1} - \varepsilon_t \text{ and } Z_{t+1} = \beta_1 + \varepsilon_{t+2} - \varepsilon_{t+1},$$

they both share a common ε_{t+1} which is highly undesirable (for future: Z_t has an MA(1) representation and is non-invertible). Similarly for the stochastic trend $Y_t = Y_{t-1} + \varepsilon_t$, taking second differences $Z_t^{(2)} = \varepsilon_t - \varepsilon_{t-1}$. Thus we encounter the same problem. Dealing with dependencies caused by over differencing induces *negative persistence* in a time series and it is a pain in the neck!

R `code`. It is straightforward to simulate a difference process. You can also use the `arima` function in R. For example, `arima.sim(list(order = c(0,1,0)), n = 200)` will simulate (2.7) and `arima.sim(list(order = c(0,2,0)), n = 200)` will simulate a differencing of order two.

Exercise 2.1 *(i) Import the yearly temperature data (file* `global_mean_temp.txt`*) into* `R` *and fit the linear model in (2.1) to the data (use the* `R` *command* `lm`, `FitTemp = lm(data), out = summary(FitTemp)`*) .*

(ii) Suppose the errors in (2.1) are correlated (linear dependence between the errors). If the errors are correlated, explain why the standard errors reported in the `R` *output may not be reliable.*

Hint: The errors are usually calculated as

$$\left(\sum_{t=1}^{n}(1,t)'(1,t)\right)^{-1}\frac{1}{n-2}\sum_{t=1}^{n}\widehat{\varepsilon}_t^2.$$

(iii) Make a plot of the residuals (over time) after fitting the linear model in (i).

(iv) Make a plot of the first differences of the temperature data (against time). Compare the plot of the residuals with the plot of the first differences.

2.3 Nonparametric methods (advanced)

In Section 2.1 we assumed that the mean had a certain known parametric form. This may not always be the case. If we have no apriori knowledge of the features in the mean, we can estimate the mean using a nonparametric approach. Of course some assumptions on the mean are still required. And the most common is to assume that the mean μ_t is a sample from a 'smooth' function. Mathematically we write that μ_t is sampled (at regular intervals) from a smooth function (i.e. u^2) with $\mu_t = \mu(\frac{t}{n})$ where the function $\mu(\cdot)$ is unknown. Under this assumption the following approaches are valid.

2.3.1 Rolling windows

Possibly one of the simplest methods is to use a 'rolling window'. There are several windows that one can use. We describe, below, the exponential window, since it can be 'evaluated' in an online way. For $t = 1$ let $\hat{\mu}_1 = Y_1$, then for $t > 1$ define

$$\widehat{\mu}_t = (1-\lambda)\widehat{\mu}_{t-1} + \lambda Y_t,$$

where $0 < \lambda < 1$. The choice of λ depends on how much weight one wants to give the present observation. The rolling window is related to the regular window often used in nonparametric regression. To see this, we note that it is straightforward to show that

$$\widehat{\mu}_t = \sum_{j=1}^{t-1}(1-\lambda)^{t-j}\lambda Y_j = \sum_{j=1}^{t}[1-\exp(-\gamma)]\exp\left[-\gamma(t-j)\right]Y_j$$

where $1-\lambda = \exp(-\gamma)$. Set $\gamma = (nb)^{-1}$ and $K(u) = \exp(-u)I(u \geq 0)$. Note that we treat n as a "sample size" (it is of the same order as n and for convenience one can let $n = t$), whereas b is a bandwidth, the smaller b the larger the weight on the current observations. Then, $\widehat{\mu}_t$ can be written as

$$\widehat{\mu}_t = \underbrace{(1-e^{-1/(nb)})}_{\approx(nb)^{-1}}\sum_{j=1}^{n}K\left(\frac{t-j}{nb}\right)Y_j,$$

where the above approximation is due to a Taylor expansion of $e^{-1/(nb)}$. This we observe that the exponential rolling window estimator is very close to a nonparametric kernel smoothing, which typically takes the form

$$\widetilde{\mu}_t = \sum_{j=1}^{n}\frac{1}{nb}K\left(\frac{t-j}{nb}\right)Y_j.$$

it is likely you came across such estimators in your nonparametric classes (a classical example is the local average where $K(u) = 1$ for $u \in [-1/2, 1/2]$ but zero elsewhere). The main difference between the rolling window estimator and the nonparametric kernel estimator is that the kernel/window for the rolling window is not symmetric. This is because we are trying to estimate the mean at time t, given only the observations up to time t. Whereas for general nonparametric kernel estimators one can use observations on both sides of t.

2.3.2 Sieve estimators

Suppose that $\{\phi_k(\cdot)\}_k$ is an orthonormal basis of $L_2[0,1]$ ($L_2[0,1] = \{f; \int_0^1 f(x)^2 dx < \infty\}$, so it includes all bounded and continuous functions)[1]. Then every function in L_2 can be represented as a linear sum of the basis. Suppose $\mu(\cdot) \in L_2[0,1]$ (for example the function is simply bounded). Then

$$\mu(u) = \sum_{k=1}^{\infty} a_k \phi_k(u).$$

Examples of basis functions are the Fourier $\phi_k(u) = \exp(iku)$, Haar/other wavelet functions etc. We observe that the unknown coefficients a_k are a linear in the 'regressors' ϕ_k. Since $\sum_k |a_k|^2 < \infty$, $a_k \to 0$. Therefore, for a sufficiently large M the finite truncation of the above is such that

$$Y_t \approx \sum_{k=1}^{M} a_k \phi_k\left(\frac{t}{n}\right) + \varepsilon_t.$$

Based on the above we observe that we can use least squares to estimate the coefficients, $\{a_k\}$. To estimate these coefficients, we truncate the above expansion to order M, and use least squares to estimate the coefficients

$$\sum_{t=1}^{n} \left[Y_t - \sum_{k=1}^{M} a_k \phi_k\left(\frac{t}{n}\right) \right]^2. \tag{2.8}$$

The orthogonality of the basis means that the corresponding design matrix $(X'X)$ is close to identity, since

$$n^{-1}(X'X)_{k_1,k_2} = \frac{1}{n} \sum_t \phi_{k_1}\left(\frac{t}{n}\right) \phi_{k_2}\left(\frac{t}{n}\right) \approx \int \phi_{k_1}(u)\phi_{k_2}(u) du = \begin{cases} 0 & k_1 \neq k_2 \\ 1 & k_1 = k_2 \end{cases}.$$

[1] Orthonormal basis means that for all k $\int_0^1 \phi_k(u)^2 du = 1$ and for any $k_1 \neq k_2$ we have $\int_0^1 \phi_{k_1}(u)\phi_{k_2}(u) du = 0$

This means that the least squares estimator of a_k is $\widehat{a}_k$ where

$$\widehat{a}_k \approx \frac{1}{n}\sum_{t=1}^{n} Y_t \phi_k\left(\frac{t}{n}\right).$$

2.4 What is trend and what is noise?

So far we have not discussed the nature of the noise ε_t. In classical statistics ε_t is usually assumed to be iid (independent, identically distributed). But if the data is observed over time, ε_t could be dependent; the previous observation influences the current observation. However, once we relax the assumption of independence in the model problems arise. By allowing the "noise" ε_t to be dependent it becomes extremely difficult to discriminate between mean trend and noise. In Figure 2.3 two plots are given. The top plot is a realisation from independent normal noise the bottom plot is a realisation from dependent noise (the AR(1) process $X_t = 0.95X_{t-1} + \varepsilon_t$). Both realisations have zero mean (no trend), but the lower plot does give the appearance of an underlying mean trend.

This effect because more problematic when analysing data where there is mean term plus dependent noise. The smoothness in the dependent noise may give the appearance of additional features mean function. This makes estimating the mean function more difficult, especially the choice of bandwidth b. To understand why, suppose the mean function is $\mu_t = \mu(\frac{t}{200})$ (the sample size $n = 200$), where $\mu(u) = 5 \times (2u - 2.5u^2) + 20$. We corrupt this quadratic function with both iid and dependent noise (the dependent noise is the AR(2) process defined in equation (2.19)). The plots are given in Figure 2.4. We observe that the dependent noise looks 'smooth' (dependence can induce smoothness in a realisation). This means that in the case that the mean has been corrupted by dependent noise it difficult to see that the underlying trend is a simple quadratic function. In a very interesting paper Hart (1991), shows that cross-validation (which is the classical method for choosing the bandwidth parameter b) is terrible when the errors are correlated.

Exercise 2.2 *The purpose of this exercise is to understand how correlated errors in a non-parametric model influence local smoothing estimators. We will use a simple local average.*

Define the smooth signal $f(u) = 5 * (2u - 2.5u^2) + 20$ *and suppose we observe* $Y_i =$

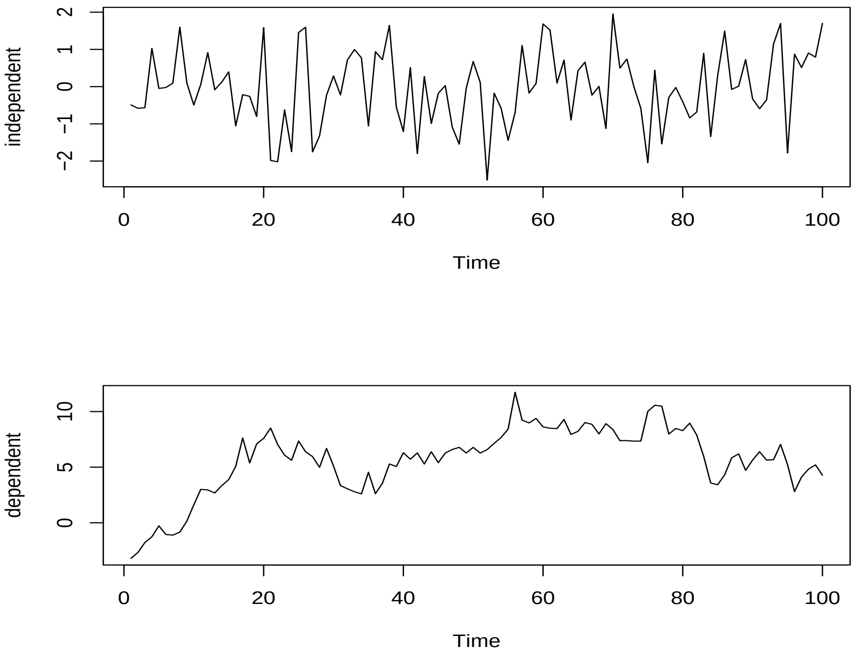


Figure 2.3: Top: realisations from iid random noise. Bottom: Realisation from dependent noise

$f(i/200) + \varepsilon_i$ *(*$n = 200$*). To simular* $f(u)$ *with* $n = 200$ *define* `temp <- c(1:200)/200` *and* `quadratic <- 5*(2*temp - 2.5*(temp**2)) + 20`*.*

(i) Simulate from the above model using iid noise. You can use the code `iid=rnom(200)` *and* `quadraticiid = (quadratic + iid)`*.*

Our aim is to estimate f*. To do this take a local average (the average can have different lengths* m*) (you can use* `mean(quadraticiid[c(k:(k+m-1))])` *for* $k = 1, \ldots, 200-m$*). Make of a plot the estimate.*

(ii) Simulate from the above model using correlated noise (we simulate from an $AR(2)$*)* `ar2 = 0.5*arima.sim(list(order=c(2,0,0), ar = c(1.5, -0.75)), n=200)` *and define* `quadraticar2 = (quadratic +ar2)`*.*

Again estimate f *using local averages and make a plot.*

Compare the plots of the estimates based on the two models above.

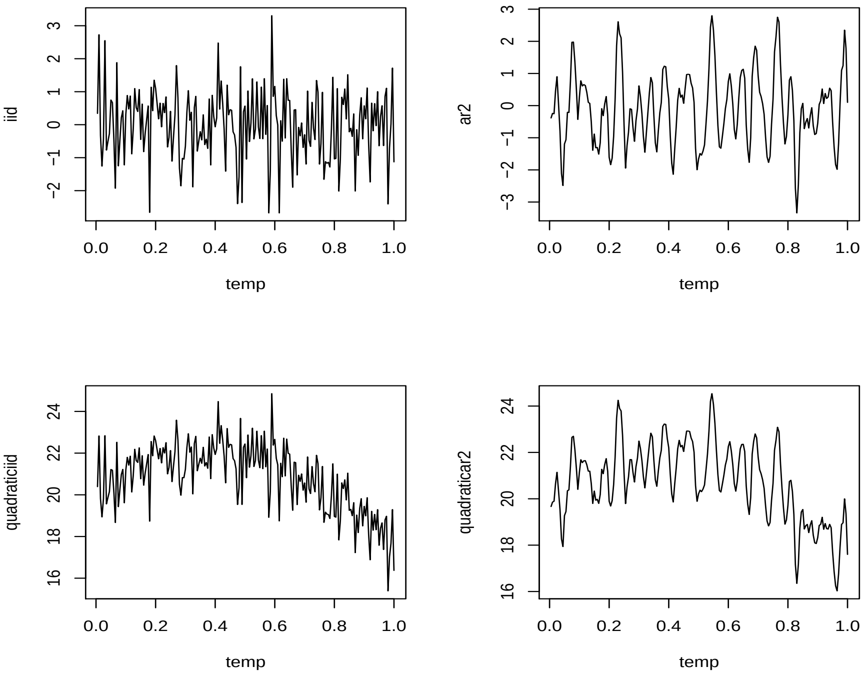


Figure 2.4: Top: realisations from iid random noise and dependent noise (left = iid and right = dependent). Bottom: Quadratic trend plus corresponding noise.

2.5 Periodic functions

Periodic mean functions arise in several applications, from ECG (which measure heart rhythms), econometric data, geostatistical data to astrostatistics. Often the aim is to estimate the period or of a periodic function. Let us return to the monthly rainfall example consider in Section 2.1, equation (2.6):

$$Y_t = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{12}\right) + \beta_3 \cos\left(\frac{2\pi t}{12}\right) + \varepsilon_t.$$

This model assumes the mean has a repetition every 12 month period. But, it assumes a very specific type of repetition over 12 months; one that is composed of one sine and one cosine. If one wanted to be more general and allow for any periodic sequence of period 12, the above should be replaced with

$$Y_t = d_{12}(t) + \varepsilon_t,$$

where $\underline{d}_{12} = (d_{12}(1), d_{12}(2), \ldots, d_{12}(12))$ and $d_{12}(t) = d_{12}(t+12)$ for all t. This a general sequence which loops every 12 time points.

In the following few sections our aim is to show that all periodic functions can be written in terms of sine and cosines.

2.5.1 The sine and cosine transform

An alternative (but equivalent) representation of this periodic sequence is by using sines and cosines. This is very reasonable, since sines and cosines are also periodic. It can be shown that

$$d_{12}(t) = a_0 + \sum_{j=1}^{5}\left[a_j \cos\left(\frac{2\pi tj}{12}\right) + b_j \sin\left(\frac{2\pi tj}{12}\right)\right] + a_6 \cos(\pi t). \tag{2.9}$$

Where we observe that the number a_j and b_js is 12, which is exactly the number of different elements in the sequence. Any periodic sequence of period 12 can be written in this way. Further equation (2.6) is the first two components in this representation. Thus the representation in (2.9) motivates why (2.6) is often used to model seasonality. You may wonder why use just the first two components in (2.9) in the seasonal, this is because typically the coefficients a_1 and b_1 are far larger than $\{a_j, b_j\}_{j=2}^{6}$. This is only a rule of thumb: generate several periodic sequences you see that in general this is true. Thus in general $\left[a_1 \cos\left(\frac{2\pi t}{12}\right) + b_1 \sin\left(\frac{2\pi t}{12}\right)\right]$ tends to capture the main periodic features in the sequence. Algebraic manipulation shows that

$$a_j = \frac{1}{12}\sum_{t=1}^{12} d_{12}(t)\cos\left(\frac{2\pi tj}{12}\right) \text{ and } b_j = \frac{1}{12}\sum_{t=1}^{12} d_{12}(t)\sin\left(\frac{2\pi tj}{12}\right). \tag{2.10}$$

These are often called the sin and cosine transforms.

In general for sequences of period P, if P is even we can write

$$d_P(t) = a_0 + \sum_{j=1}^{P/2-1}\left[a_j \cos\left(\frac{2\pi tj}{P}\right) + b_j \sin\left(\frac{2\pi tj}{P}\right)\right] + a_{P/2}\cos(\pi t) \tag{2.11}$$

and if P is odd

$$d_P(t) = a_0 + \sum_{j=1}^{\lfloor P/2 \rfloor - 1} \left[a_j \cos\left(\frac{2\pi tj}{P}\right) + b_j \sin\left(\frac{2\pi tj}{P}\right) \right] \tag{2.12}$$

where

$$a_j = \frac{1}{P} \sum_{t=1}^{P} d_P(t) \cos\left(\frac{2\pi tj}{P}\right) \text{ and } b_j = \frac{1}{P} \sum_{t=1}^{P} d_P(t) \sin\left(\frac{2\pi tj}{P}\right).$$

The above reconstructs the periodic sequence $d_P(t)$ in terms of sines and cosines. What we will learn later on is that all sequences can be built up with sines and cosines (it does not matter if they are periodic or not).

2.5.2 The Fourier transform (the sine and cosine transform in disguise)

We will now introduce a tool that often invokes panic in students. But it is very useful and is simply an alternative representation of the sine and cosine transform (which does not invoke panic). If you tried to prove (2.10) you would have probably used several cosine and sine identities. It is a very mess proof. A simpler method is to use an alternative representation which combines the sine and cosine transforms and imaginary numbers. We recall the identity

$$e^{i\omega} = \cos(\omega) + i\sin(\omega).$$

where $i = \sqrt{-1}$. $e^{i\omega}$ contains the sin and cosine information in just one function. Thus $\cos(\omega) = \operatorname{Re} e^{i\omega} = (e^{i\omega} + e^{-i\omega})/2$ and $\sin(\omega) = \operatorname{Im} e^{i\omega} = -i(e^{i\omega} - e^{-i\omega})/2$.

It has some very useful properties that just require basic knowledge of geometric series. We state these below. Define the ratio $\omega_{k,n} = 2\pi k/n$ (we exchange 12 for n), then

$$\sum_{k=0}^{n-1} \exp(ij\omega_{k,n}) = \sum_{k=0}^{n-1} \exp(ik\omega_{j,n}) = \sum_{k=0}^{n-1} [\exp(i\omega_{j,n})]^k.$$

Keep in mind that $j\omega_{k,n} = j2\pi k/n = k\omega_{j,n}$. If $j = 0$, then $\sum_{k=0}^{n-1} \exp(ij\omega_{n,n}) = n$. On the other hand, if $1 \leq j, k \leq (n-1)$, then $\exp(ij\omega_{k,n}) = \cos(2j\pi k/n) + i\sin(2j\pi k/n) \neq 1$. And we can use the geometric sum identity

$$\sum_{k=0}^{n-1} \exp(ij\omega_{k,n}) = \sum_{k=0}^{n-1} [\exp(i\omega_{j,n})]^k = \frac{1 - \exp(in\omega_{k,n})}{1 - \exp(i\omega_{k,n})}.$$

But $\exp(in\omega_{k,n}) = \cos(n2\pi k/n) + i\sin(n2\pi k/n) = 1$. Thus for $1 \leq k \leq (n-1)$ we have

$$\sum_{k=0}^{n-1} \exp(ij\omega_{k,n}) = \frac{1 - \exp(in\omega_{j,n})}{1 - \exp(i\omega_{j,n})} = 0.$$

In summary,

$$\sum_{k=0}^{n-1} \exp(ij\omega_{k,n}) = \begin{cases} n & j = n \text{ or } 0 \\ 0 & 1 \leq j \leq (n-1) \end{cases} \tag{2.13}$$

Now using the above results we now show we can rewrite $d_{12}(t)$ in terms of $\exp(i\omega)$ (rather than sines and cosines). And this representation is a lot easier to show; though you it is in terms of complex numbers. Set $n = 12$ and define the coefficient

$$A_{12}(j) = \frac{1}{12} \sum_{t=0}^{11} d_{12}(t) \exp\left(it\omega_{j,12}\right).$$

$A_{12}(j)$ is complex (it has real and imaginary parts), with a little thought you can see that $A_{12}(j) = \overline{A_{12}(12-j)}$. By using (2.13) it is easily shown (see below for proof) that

$$d_{12}(\tau) = \sum_{j=0}^{11} A_{12}(j) \exp(-ij\omega_{\tau,12}) \tag{2.14}$$

This is just like the sine and cosine representation

$$d_{12}(t) = a_0 + \sum_{j=1}^{5} \left[a_j \cos\left(\frac{2\pi tj}{12}\right) + b_j \sin\left(\frac{2\pi tj}{12}\right) \right] + a_6 \cos\left(\pi t\right).$$

but with $\exp(ij\omega_{t,12})$ replacing $\cos(j\omega_{t,12})$ and $\sin(j\omega_{t,12})$.

Proof of equation (2.14) The proof of (2.14) is very simple and we now give it. Plugging in the equation for $A_{12}(j)$ into (2.14) gives

$$\begin{aligned} d_{12}(\tau) = \sum_{j=0}^{11} A_{12}(j)\exp(-ij\omega_{\tau,12}) &= \frac{1}{12}\sum_{t=0}^{11} d_{12}(t)\sum_{j=0}^{11}\exp\left(it\omega_{j,n}\right)\exp(-ij\omega_{\tau,12}) \\ &= \frac{1}{12}\sum_{t=0}^{11} d_{12}(t)\sum_{j=0}^{11}\exp\left(i(t-\tau)\omega_{j,12})\right). \end{aligned}$$

We know from (2.13) that $\sum_{j=0}^{11}\exp\left(i(t-\tau)\omega_{j,12})\right) = 0$ unless $t = \tau$. If $t = \tau$, then $\sum_{j=0}^{11}\exp\left(i(t-\tau)\omega_{j,12})\right) = 12$. Thus

$$\begin{aligned} \frac{1}{12}\sum_{t=0}^{11} d_{12}(t)\sum_{j=0}^{11}\exp\left(i(t-\tau)\omega_{j,12})\right) &= \frac{1}{12}\sum_{t=0}^{11} d_{12}(t)I(t=\tau)\times 12 \\ &= d_{12}(t), \end{aligned}$$

this proves (2.14). □

Remember the above is just writing the sequence in terms of its sine and cosine transforms in fact it is simple to link the two sets of coefficients:

$$\begin{aligned} a_j &= \operatorname{Re} A_{12}(j) = \frac{1}{2}\left[A_{12}(j) + A_{12}(12-j)\right] \\ b_j &= \operatorname{Im} A_{12}(j) = \frac{-i}{2}\left[A_{12}(j) - A_{12}(12-j)\right]. \end{aligned}$$

We give an example of a periodic function and its Fourier coefficients (real and imaginary parts) in Figure 2.5. The peak at the zero frequency of the real part corresponds to the mean of the periodic signal (if the mean is zero, this will be zero).

Example 2.5.1 *In the case that $d_P(t)$ is a pure sine or cosine function* $\sin(2\pi t/P)$ *or* $\cos(2\pi t/P)$, *then $A_P(j)$ will only be non-zero at $j = 1$ and $j = P - 1$.*

This is straightfoward to see, but we formally prove it below. Suppose that $d_P(t) =$

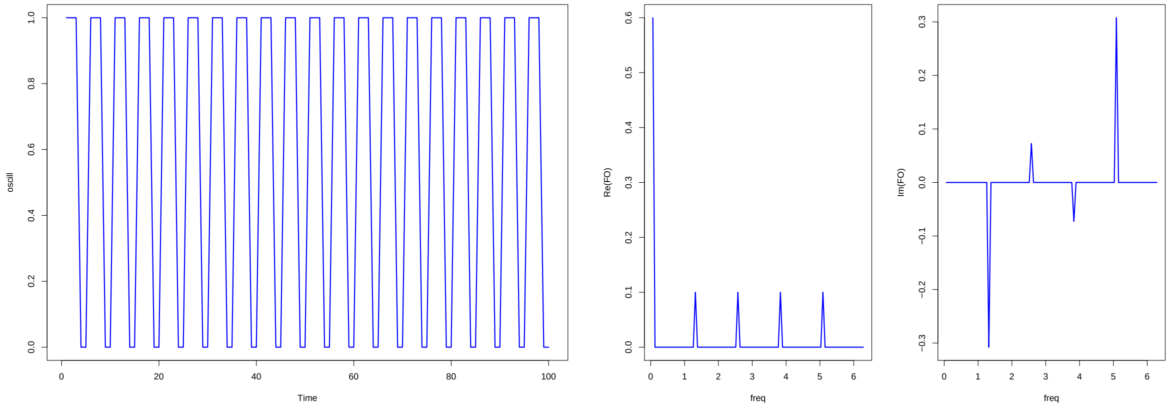


Figure 2.5: Left: Periodic function $d_5(s) = 1$ for $s = 1, 2$, $d_5(s) = 0$ for $s = 3, 4, 5$ (period 5), Right: The real and imaginary parts of its Fourier transform

$\cos\left(\frac{2\pi s}{P}\right)$, *then*

$$\frac{1}{P}\sum_{s=0}^{P-1}\cos\left(\frac{2\pi s}{P}\right)\exp\left(i\frac{2\pi sj}{P}\right) = \frac{1}{2P}\sum_{s=0}^{P-1}\left(e^{i2\pi s/P} + e^{-i2\pi s/P}\right)e^{i\frac{2\pi sj}{P}} = \begin{cases} 1/2 & j = 1 \text{ or } P-1 \\ 0 & \text{otherwise} \end{cases}$$

Suppose that $d_P(t) = \sin\left(\frac{2\pi s}{P}\right)$, *then*

$$\frac{1}{P}\sum_{s=0}^{P-1}\sin\left(\frac{2\pi s}{P}\right)\exp\left(i\frac{2\pi sj}{P}\right) = \frac{-i}{2P}\sum_{s=0}^{P-1}\left(e^{i2\pi s/P} - e^{-i2\pi s/P}\right)e^{i\frac{2\pi sj}{P}} = \begin{cases} i/2 & j = 1 \\ -i/2 & j = P-1 \\ 0 & \text{otherwise} \end{cases}$$

2.5.3 The discrete Fourier transform

The discussion above shows that any periodic sequence can be written as the sum of (modulated) sins and cosines up to that frequency. But the same is true for any sequence. Suppose $\{Y_t\}_{t=1}^{n}$ is a sequence of length n, then it can always be represented as the superposition of n sine and cosine functions. To make calculations easier we use $\exp(ij\omega_{k,n})$ instead of sines and cosines:

$$Y_t = \sum_{j=0}^{n-1} A_n(j)\exp(-it\omega_{j,n}), \qquad (2.15)$$

where the amplitude $A_n(j)$ is

$$A_n(j) = \frac{1}{n}\sum_{\tau=1}^{n} Y_\tau \exp(i\tau\omega_{j,n}).$$

Here Y_t is acting like $d_P(t)$, it is also periodic if we over the boundary $[1,\ldots,n]$. By using (2.15) as the definition of Y_t we can show that $Y_{t+n} = Y_t$.

Often the n is distributed evenly over the two sums and we represent Y_t as

$$Y_t = \frac{1}{\sqrt{n}}\sum_{k=0}^{n-1} J_n(\omega_{k,n})\exp(-it\omega_{k,n}),$$

where the amplitude of $\exp(-it\omega_{k,n})$ is

$$J_n(\omega_{k,n}) = \frac{1}{\sqrt{n}}\sum_{\tau=1}^{n} Y_\tau \exp(i\tau\omega_{k,n}).$$

This representation evenly distributes $1/\sqrt{n}$ amongst the two sums. $J_n(\omega_{k,n})$ is called the Discrete Fourier transform (DFT) of $\{Y_t\}$. It serves a few purposes:

- $J_n(\omega_{k,n})$ measures the contribution (amplitude) of $\exp(it\omega_{k,n})$ (or $\cos(t\omega_{k,n})$ and $\sin(t\omega_{k,n})$) in $\{Y_t\}$.
- $J_n(\omega_{k,n})$ is a linear transformation of $\{Y_t\}_{t=1}^{n}$.
- You can view $J_n(\omega_{k,n})$ as a scalar product of $\{Y_t\}$ with sines and cosines, or as projection onto sines or cosines or measuring the resonance of $\{Y_t\}$ at frequency $\omega_{k,n}$. It has the benefit of being a microscope for detecting periods, as we will see in the next section.

For general time series, the DFT, $\{J_n(\frac{2\pi k}{n}); 1 \leq k \leq n\}$ is simply a decomposition of the time series $\{X_t; t = 1,\ldots,n\}$ into sins and cosines of different frequencies. The magnitude of $J_n(\omega_k)$ informs on how much of the functions $\sin(t\omega)$ and $\cos(t\omega_k)$ are in the $\{X_t; t = 1,\ldots,n\}$. Below we define the periodogram. The periodogram effectively removes the complex part in $J_n(\omega_k)$ and only measures the absolute magnitude.

Definition 2.5.1 (The periodogram) *$J_n(\omega)$ is complex random variables. Often the absolute square of $J_n(\omega)$ is analyzed, this is called the periodogram*

$$I_n(\omega) = |J_n(\omega)|^2 = \frac{1}{n}\left|\sum_{t=1}^{n} X_t \cos(t\omega)\right|^2 + \frac{1}{n}\left|\sum_{t=1}^{n} X_t \sin(t\omega)\right|^2.$$

$I_n(\omega)$ combines the information in the real and imaginary parts of $J_n(\omega)$ and has the advantage that it is real.

$I_n(\omega)$ is symmetric about π. It is also periodic every $[0, 2\pi]$, thus $I_n(\omega + 2\pi) = I_n(\omega)$.

Put together only needs to consider $I_n(\omega)$ in the range $[0, \pi]$ to extract all the information from $I_n(\omega)$.

2.5.4 The discrete Fourier transform and periodic signals

In this section we consider signals with periodic trend:

$$\begin{aligned} Y_t &= d_P(t) + \varepsilon_t \qquad t = 1, \ldots, n \\ &= \sum_{j=0}^{P-1} A_P(j) e^{-i\frac{2\pi jt}{P}} + \varepsilon_t \end{aligned}$$

where for all t, $d_P(t) = d_P(t + P)$ (assume $\{\varepsilon_t\}$ are iid). Our aim in this section is estimate (at least visually) the period. We use the DFT of the time series to gain some standing of $d_P(t)$. We show below that the linear transformation $J_n(\omega_{k,n})$ is more informative about d_P that $\{Y_t\}$.

We recall that the discrete Fourier transform of $\{Y_t\}$ is

$$J_n(\omega_{k,n}) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} Y_t \left[\cos(t\omega_{k.n}) - i\sin(t\omega_k)\right] = \sum_{t=1}^{n} Y_t \exp(-it\omega_{k,n})$$

where $\{\omega_k = \frac{2\pi k}{n}\}$. We show below that when the periodicity in the cosine and sin function matches the periodicity of the mean function $J_n(\omega)$ will be large and at other frequencies it will be small. Thus

$$J_n(\omega_{k,n}) \quad = \quad \begin{cases} \sqrt{n}A_p(r) + \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t e^{-it\omega_{k,n}} & k = \frac{n}{P}r, \quad r = 0,\ldots,P-1. \\ \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t e^{-it\omega_{k,n}} & k \neq \frac{n}{P}\mathbb{Z} \end{cases} \tag{2.16}$$

Assuming that $\sum_{t=1}^{n}\varepsilon_t e^{-it\omega_{k,n}}$ is low lying noise (we discuss this in detail later), what we should see are P large spikes, each corresponding to $A_P(r)$. Though the above is simply an algebraic calculation. The reason for the term n in (2.16) (recall n is the sample size) is because there are n/P repetitions of the period.

Example We consider a simple example where $d_4(s) = (1.125, -0.375, -0.375, -0.375)$ (period $= 4$, total length 100, number of repetitions 25). We add noise to it (iid normal with $\sigma = 0.4$). A plot of one realisation is given in Figure 2.7. In Figure 2.8 we superimpose the observed signal with with two different sine functions. Observe that when the sine function matches the frequencies ($\sin(25u)$, red plot) their scalar product will be large. But when the sin frequency does not match the periodic frequency the scalar product will be close to zero. In

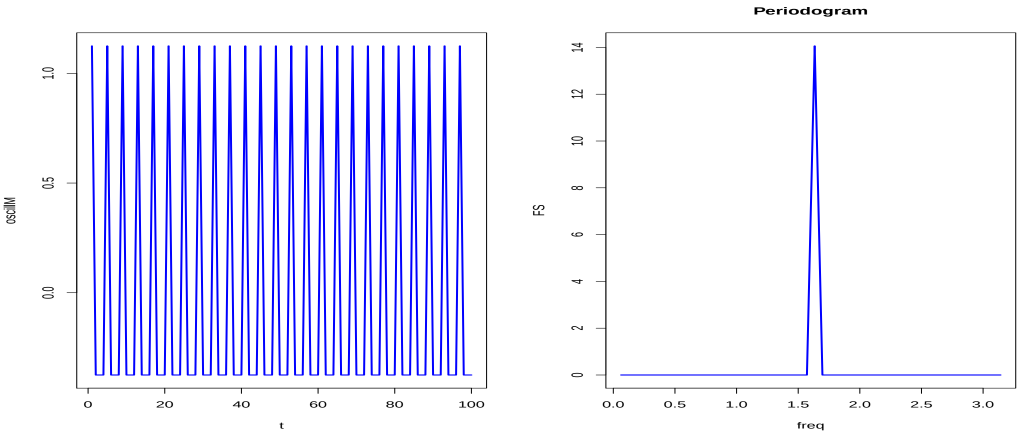


Figure 2.6: Left: Periodic function $d_4(s) = (1.125, -0.375, -0.375, -0.375)$ (period 4)

In Figure 2.9 we plot the signal together with is periodgram. Observe that the plot matches equation (2.16). At the frequency of the period the signal amplitude is very large.

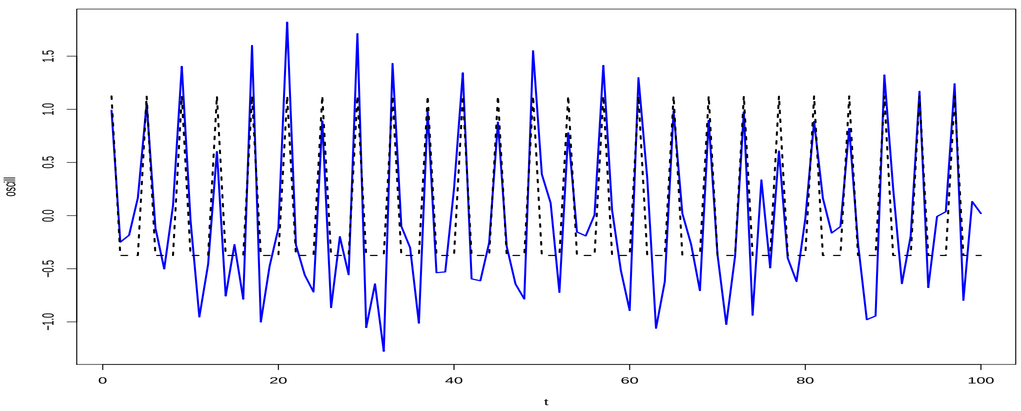


Figure 2.7: Periodic function $d_4(s) = (1.125, -0.375, -0.375, -0.375)$ (period 4) and signal with noise (blue line).

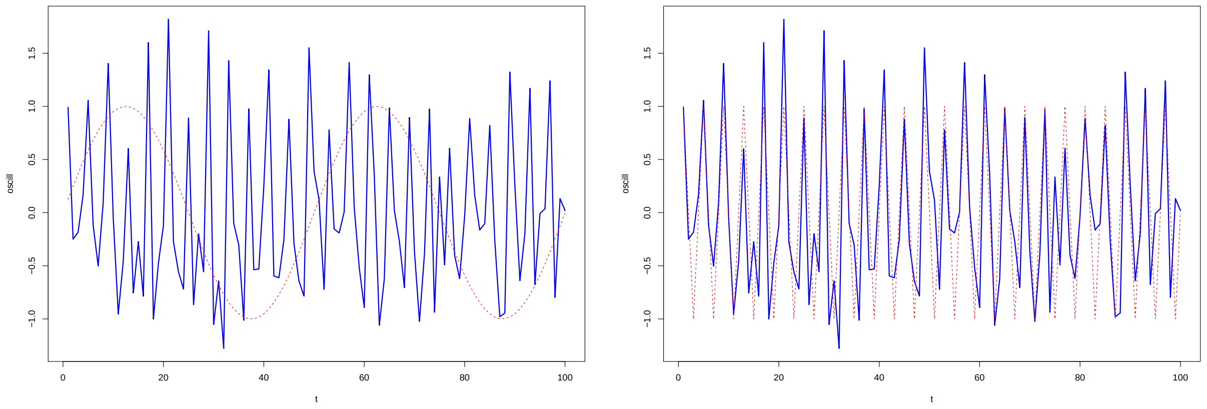


Figure 2.8: Left: Signal superimposed with $\sin(u)$. Right: Signal superimposed with $\sin(25u)$.

Proof of equation (2.16) To see why, we rewrite $J_n(\omega_k)$ (we assume n is a multiple of P) as

$$
\begin{aligned}
J_n(\omega_k) &= \frac{1}{\sqrt{n}}\sum_{t=0}^{n} d_P(t)\exp(it\omega_k) + \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t e^{it\omega_k} \\
&= \frac{1}{\sqrt{n}}\sum_{t=0}^{n/P-1}\sum_{s=1}^{P} d_P(Pt+s)\exp(iPt\omega_k + is\omega_k) + \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t e^{it\omega_k} \\
&= \frac{1}{\sqrt{n}}\sum_{t=0}^{n/P-1}\exp(iPt\omega_k)\sum_{s=1}^{P} d_P(s)\exp(is\omega_k) + \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t e^{it\omega_k} \\
&= \frac{1}{\sqrt{n}}\sum_{s=1}^{P} d_P(s)\exp(is\omega_k)\sum_{t=0}^{n/P-1}\exp(iPt\omega_k) + \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t e^{it\omega_k}.
\end{aligned}
$$

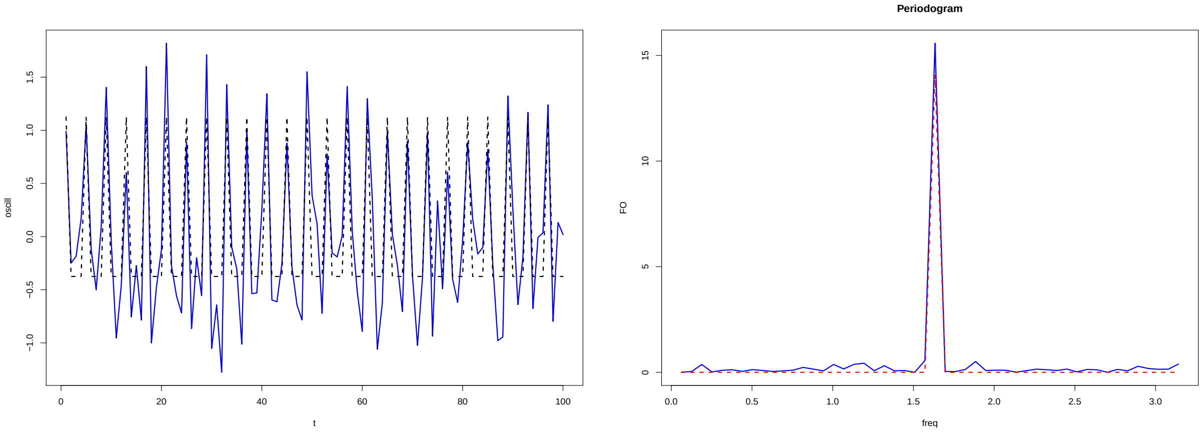


Figure 2.9: Left: Signal, Right: periodogram of signal (peridogram of periodic function in red)

We now use a result analogous to (2.13)

$$\sum_{t=0}^{n/P-1}\exp(iPt\omega_k) = \begin{cases} \frac{\exp(i2\pi k)}{1-\exp(iPt\omega_k)} = 0 & k \neq \frac{n}{P}\mathbb{Z} \\ n/P & k \in \frac{n}{P}\mathbb{Z} \end{cases}$$

Thus

$$J_n(\omega_k) \quad = \quad \begin{cases} \sqrt{n}A_p(r) + \sum_{t=1}^{n}\varepsilon_t e^{it\omega_k} & k = \frac{n}{P}r, \quad r = 0,\ldots,P-1. \\ \sum_{t=1}^{n}\varepsilon_t e^{it\omega_k} & k \neq \frac{n}{P}\mathbb{Z} \end{cases}$$

where $A_P(r) = P^{-1}\sum_{s=1}^{P} d_P(s)\exp(2\pi isr/P)$. This proves (2.16) □

Exercise 2.3 *Generate your own periodic sequence of length P (you select P). Call this sequence $\{d_P(t)\}$ and generate a sequence $\{x_t\}$ with several replications of $\{d_P(t)\}$ and calculate the periodogram of the periodic signal.*

Add iid noise to the signal and again evaluate the periodogram (do the same for noise with different standard deviations).

(i) Make plots of the true signal and the corrupted signal.

(i) Compare the periodogram of the true signal with the periodogram of the corrupted signal.

2.5.5 Smooth trends and its corresponding DFT

So far we have used the DFT to search for periodocities. But the DFT/periodogram of a smooth signal also leaves an interesting signature. Consider the quadratic signal

$$g(t) = 6\left[\frac{t}{100} - \left(\frac{t}{100}\right)^2\right] - 0.7 \qquad t = 1, \ldots, 100.$$

To $g(t)$ we add iid noise $Y_t = g(t) + \varepsilon_t$ where $\mathrm{var}[\varepsilon_t] = 0.5^2$. A realisation and its corresponding periodogram is given in Figure 2.10. We observe that the quadratic signal is composed of low frequencies (sines and cosines with very large periods). In general, any signal which is "smooth" can be decomposed of sines and cosines in the very low frequencies. Thus a periodogram with a large peak around the low frequencies, suggests that the underlying signal contains a smooth signal (either deterministically or stochastically).

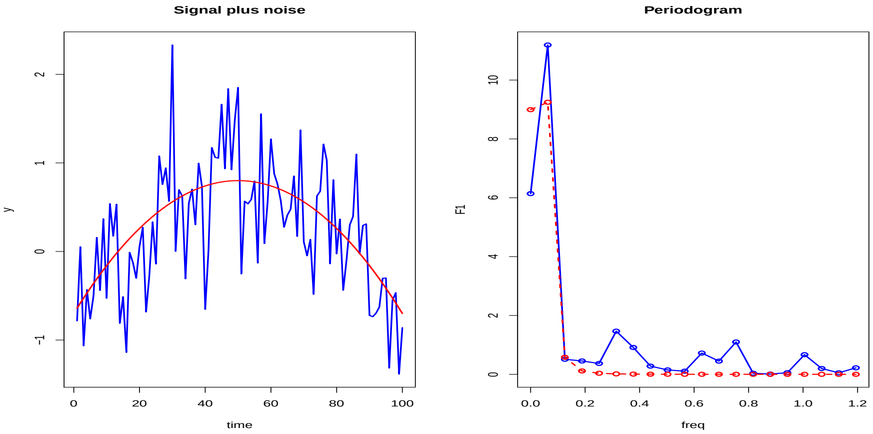


Figure 2.10: Left: Signal and noise (blue). The signal is in red. Right: Periodogram of signal plus noise (up to frequency $\pi/5$). Periodogram of signal is in red.

2.5.6 Period detection

In this section we formalize what we have seen and derived for the periodic sequences given above. Our aim is to estimate the period P. But to simplify the approach, we focus on the case that $d_P(t)$ is a pure sine or cosine function (no mix of sines and cosines).

We will show that the visual Fourier transform method described above is equivalent to period estimation using least squares. Suppose that the observations $\{Y_t; t = 1, \ldots, n\}$

satisfy the following regression model

$$Y_t = A\cos(\Omega t) + B\sin(\Omega t) + \varepsilon_t = A\cos\left(\frac{2\pi t}{P}\right) + B\sin\left(\frac{2\pi t}{P}\right) + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid standard normal random variables and $0 < \Omega < \pi$ (using the periodic notation we set $\Omega = \frac{2\pi}{P}$).

The parameters A, B, and Ω are real and unknown. Unlike the regression models given in (2.1) the model here is nonlinear, since the unknown parameter, Ω, is inside a trignometric function. Standard least squares methods cannot be used to estimate the parameters. Assuming Gaussianity of $\{\varepsilon_t\}$ (though this assumption is not necessary), the maximum likelihood corresponding to the model is

$$\mathcal{L}_n(A, B, \Omega) = -\frac{1}{2}\sum_{t=1}^{n}(Y_t - A\cos(\Omega t) - B\sin(\Omega t))^2$$

(alternatively one can think of it in terms use least squares which is negative of the above). The above criterion is a negative nonlinear least squares criterion in A, B and Ω. It does not yield an analytic solution and would require the use of a numerical maximisation scheme. However, using some algebraic manipulations, explicit expressions for the estimators can be obtained (see Walker (1971) and Exercise 2.5). The result of these manipulations give the frequency estimator

$$\widehat{\Omega}_n = \arg\max_{\omega} I_n(\omega)$$

where

$$I_n(\omega) = \frac{1}{n}\left|\sum_{t=1}^{n} Y_t \exp(it\omega)\right|^2 = \frac{1}{n}\left(\sum_{t=1}^{n} Y_t\cos(t\Omega)\right)^2 + \frac{1}{n}\left(\sum_{t=1}^{n} Y_t\sin(t\omega)\right)^2. \qquad (2.17)$$

Using $\widehat{\Omega}_n$ we estimate A and B with

$$\widehat{A}_n = \frac{2}{n}\sum_{t=1}^{n} Y_t\cos(\widehat{\Omega}_n t) \text{ and } \widehat{B}_n = \frac{2}{n}\sum_{t=1}^{n} Y_t\sin(\widehat{\Omega}_n t).$$

The rather remarkable aspect of this result is that the rate of convergence of

$$|\widehat{\Omega}_n - \Omega| = O_p(n^{-3/2}),$$

which is faster than the standard $O(n^{-1/2})$ that we usually encounter (we will see this in Example 2.5.2). This means that for even moderate sample sizes if $P = \frac{2\pi}{\Omega}$ is not too large, then $\widehat{\Omega}_n$ will be "close" to Ω. [2]. The reason we get this remarkable result was alluded to previously. We reiterate it again

$$I_n(\omega) \approx \underbrace{\frac{1}{n}\left|\sum_{t=1}^{n}\left[A\cos(t\Omega) + B\sin(t\Omega)\right]e^{it\omega}\right|}_{\text{signal}} + \underbrace{\frac{1}{n}\left|\sum_{t=1}^{n}\varepsilon_t e^{it\omega}\right|^2}_{\text{noise}}.$$

The "signal" in $I_n(\omega_k)$ is the periodogram corresponding to the cos and/or sine function. For example setting $\Omega = 2\pi/P$, $A = 1$ and $B = 0$. The signal is

$$\frac{1}{n}\left|\sum_{t=1}^{n}\cos\left(\frac{2\pi t}{P}\right)e^{it\omega_k}\right|^2 = \left\{\begin{array}{cc} \frac{n}{4} & k = \frac{n}{P} \text{ or } k = \frac{n-P}{P} \\ 0 & \text{other wise} \end{array}\right. .$$

Observe there is a peak at $\frac{2\pi P}{n}$ and $\frac{2\pi(n-P)}{n}$, which is of size n, elsewhere it is zero. On the other hand the noise is

$$\frac{1}{n}\left|\sum_{t=1}^{n}\varepsilon_t e^{it\omega_k}\right|^2 = \left|\underbrace{\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t e^{it\omega_k}}_{\text{treat as a rescaled mean}}\right|^2 = O_p(1),$$

where $O_p(1)$ means that it is bounded in probability (it does not grow as $n \to \infty$). Putting these two facts together, we observe that the contribution of the signal dominates the periodogram $I_n(\omega)$. A simulation to illustrate this effect is given in Figure **??**

Remark 2.5.1 *In practice, usually we evaluate $J_n(\omega)$ and $I_n(\omega)$ at the so called fundamental*

[2] In contrast consider the iid random variables $\{X_t\}_{t=1}^n$, where $\mathrm{E}[X_t] = \mu$ and $\mathrm{var}(X_t) = \sigma^2$. The variance of the sample mean $\bar{X} = n^{-1}\sum_{t=1}^n$ is $\mathrm{var}[\bar{X}] = \sigma^2/n$ (where $\mathrm{var}(X_t) = \sigma^2$). This means $|\bar{X} - \mu| = O_p(n^{-1/2})$. This means there exists a random variable U such that $|\bar{X} - \mu| \leq n^{-1/2}U$. Roughly, this means as $n \to \infty$ the distance between $\bar{X}$ and μ declines at the rate $n^{-1/2}$.

frequencies $\omega_k = \frac{2\pi k}{n}$ *and we do this with the* `fft` *function in R:*

$$\{Y_t\}_{t=1}^n \to \left\{ J_n(\frac{2\pi k}{n}) = \frac{1}{\sqrt{n}}\sum_{t=1}^n Y_t \cos\left(t\frac{2\pi k}{n}\right) + i\frac{1}{\sqrt{n}}\sum_{t=1}^n Y_t \sin\left(t\frac{2\pi k}{n}\right)\right\}_{k=1}^n .$$

$J_n(\omega_k)$ *is simply a* linear *one to one transformation of the data (nothing is lost in this transformation). Statistical analysis can be applied on any transformation of the data (for example Wavelet transforms). It so happens that for stationary time series this so called Fourier transform has some advantages.*

For period detection and amplitude estimation one can often obtain a better estimator of P *(or* Ω*) if a finer frequency resolution were used. This is done by padding the signal with zeros and evaluating the periodogram on* $\frac{2\pi k}{d}$ *where* $d >> n$*. The estimate of the period is then evaluated by using*

$$\widehat{P} = \frac{d}{\widehat{K}-1}$$

where $\widehat{K}$ *is the entry in the vector corresponding to the maximum of the periodogram.*

We consider an example below.

Example 2.5.2 *Consider the following model*

$$Y_t = 2\sin\left(\frac{2\pi t}{8}\right) + \varepsilon_t \qquad t = 1,\ldots,n. \tag{2.18}$$

where ε_t *are iid standard normal random variables (and for simplicity we assume* n *is a multiple of* 8*). Note by using Remark 2.5.1 and equation (2.16) we have*

$$\frac{1}{n}\left|2\sum_{t=1}^n \sin\left(\frac{2\pi t}{8}\right)\exp(it\omega_{k,n})\right|^2 = \left\{ \begin{array}{ll} n & k = \frac{n}{8} \textit{ or } n - \frac{n}{8} \\ 0 & \textit{otherwise} \end{array}\right.$$

It is clear that $\{Y_t\}$ *is made up of a periodic signal with period eight. We make a plot of one realisation (using sample size* $n = 128$*) together with the periodogram* $I(\omega)$ *(defined in (2.17)). In Figure 2.11 we give a plot of one realisation together with a plot of the* *periodogram. From the realisation, it is not clear what the period is (the noise has made it difficult to see the period). On the other hand, the periodogram clearly shows a peak at frequenct $2\pi/8 \approx 0.78$ (where we recall that 8 is the period) and $2\pi - 2\pi/8$ (since the periodogram is symmetric about π).*

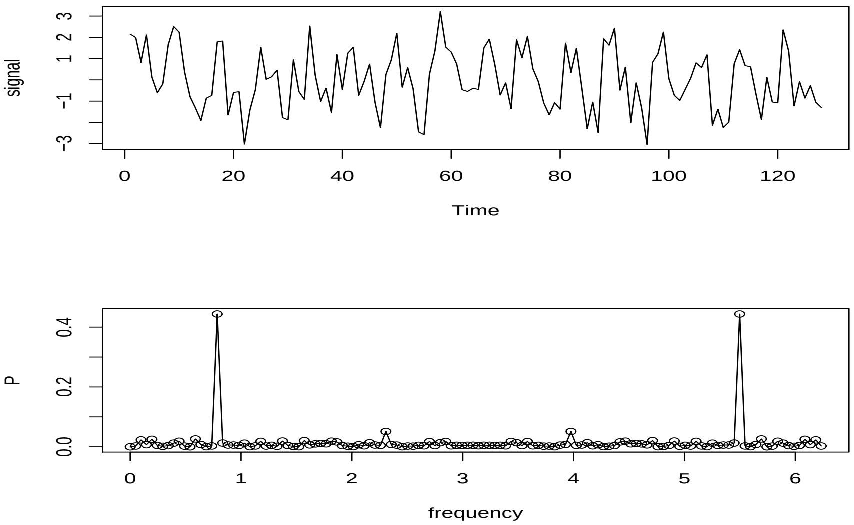


Figure 2.11: Left: Realisation of (2.18) plus iid noise, Right: Periodogram of signal plus iid noise.

Searching for peaks in the periodogram is a long established method for detecting periodicities. The method outlined above can easily be generalized to the case that there are multiple periods. However, distinguishing between two periods which are very close in frequency (such data arises in astronomy) is a difficult problem and requires more subtle methods (see Quinn and Hannan (2001)).

The Fisher's g-statistic (advanced) The discussion above motivates Fisher's test for hidden period, where the objective is to detect a period in the signal. The null hypothesis is H_0 : The signal is just white noise with no periodicities the alternative is H_1 : The signal contains a periodicity. The original test statistic was constructed under the assumption that the noise was iid Gaussian. As we have discussed above, if a period exists, $I_n(\omega_k)$ will contain a few "large" values, which correspond to the periodicities. The majority of $I_n(\omega_k)$ will be "small".

Based on this notion, the Fisher's g-statistic is defined as

$$\eta_n = \frac{\max_{1\leq k\leq (n-1)/2} I_n(\omega_k)}{\frac{2}{n-1}\sum_{k=1}^{(n-1)/2} I_n(\omega_k)},$$

where we note that the denominator can be treated as the average noise. Under the null (and iid normality of the noise), this ratio is pivotal (it does not depend on any unknown nuisance parameters).

2.5.7 Period detection and correlated noise

The methods described in the previous section are extremely effective if the error process $\{\varepsilon_t\}$ is uncorrelated. However, problems arise when the errors are correlated. To illustrate this issue, consider again model (2.18)

$$Y_t = 2\sin\left(\frac{2\pi t}{8}\right) + \varepsilon_t \qquad t = 1,\ldots,n.$$

but this time the errors are correlated. More precisely, they are generated by the AR(2) model,

$$\varepsilon_t = 1.5\varepsilon_{t-1} - 0.75\varepsilon_{t-2} + \epsilon_t, \tag{2.19}$$

where $\{\epsilon_t\}$ are iid random variables (do not worry if this does not make sense to you we define this class of models precisely in Chapter 4). As in the iid case we use a sample size $n = 128$. In Figure 2.12 we give a plot of one realisation and the corresponding periodogram. We observe that the peak at $2\pi/8$ is not the highest. The correlated errors (often called coloured noise) is masking the peak by introducing new peaks. To see what happens for larger sample sizes, we consider exactly the same model (2.18) with the noise generated as in (2.19). But this time we use $n = 1024$ (8 time the previous sample size). A plot of one realisation, together with the periodogram is given in Figure 2.13. In contrast to the smaller sample size, a large peak is visible at $2\pi/8$. These examples illustrates two important points:

(i) When the noise is correlated and the sample size is relatively small it is difficult to

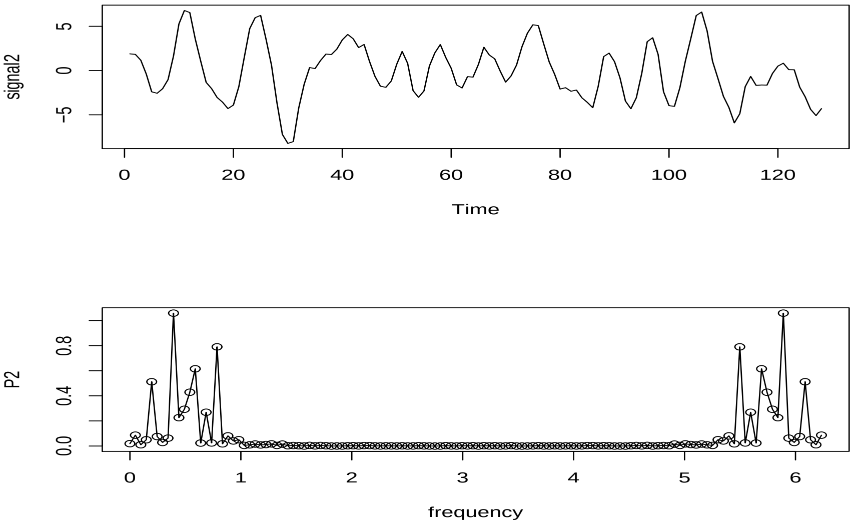


Figure 2.12: Top: Realisation of (2.18) plus correlated noise and $n = 128$, Bottom: Periodogram of signal plus correlated noise.

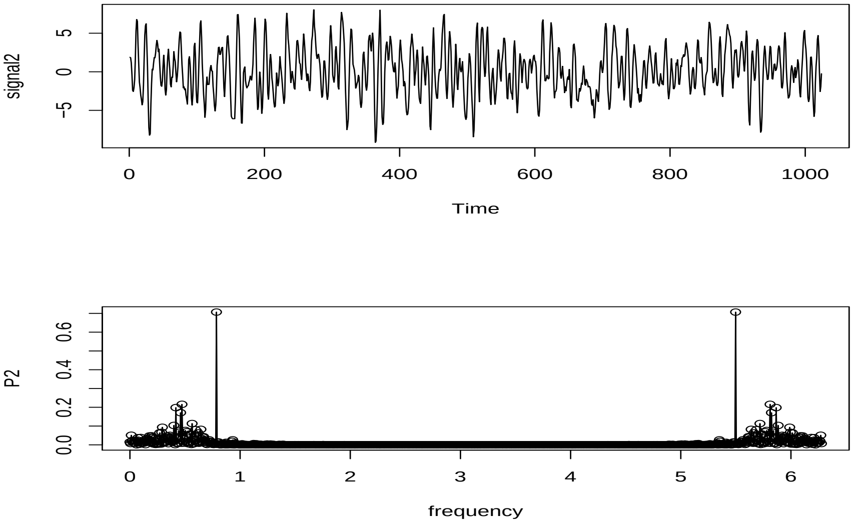


Figure 2.13: Top: Realisation of (2.18) plus correlated noise and $n = 1024$, Bottom: Periodogram of signal plus correlated noise.

disentangle the deterministic period from the noise. Indeed we will show in Chapters 4 and 6 that linear time series (such as the AR(2) model described in (2.19)) can exhibit similar types of behaviour to a periodic deterministic signal. This is a subject of on going research that dates back at least 60 years (see Quinn and Hannan (2001) and the P-statistic proposed by Priestley).

However, the similarity is only to a point. Given a large enough sample size (which may in practice not be realistic), the deterministic frequency dominates again (as we have seen when we increase n to 1024).

(ii) The periodogram holds important information about oscillations in the both the signal and also the noise $\{\varepsilon_t\}$. If the noise is iid then the corresponding periodogram tends to be flatish (see Figure 2.11). This informs us that no frequency dominates others. And is the reason that iid time series (or more precisely uncorrelated time series) is called "white noise".

Comparing Figure 2.11 with 2.12 and 2.13) we observe that the periodogram does not appear completely flat. Some frequencies tend to be far larger than others. This is because when data is dependent, certain patterns are seen, which are registered by the periodogram (see Section 4.3.6).

Understanding the DFT and the periodogram is called spectral analysis and is explored in Chapters 10 and 11.

2.5.8 History of the periodogram

The use of the periodogram, $I_n(\omega)$ to detect for periodocities in the data dates back to Schuster in the 1890's. One of Schuster's interest was sunspot data. He analyzed the number of sunspot through the lense of the periodogram. A plot of the monthly time series and corresponding periodogram is given in Figure 2.14. Let $\{Y_t\}$ denote the number of sunspots at month t. Schuster fitted a model of the type the period trend plus noise model

$$Y_t = A\cos(\Omega t) + B\sin(\Omega t) + \varepsilon_t,$$

$\Omega = 2\pi/P$. The periodogram below shows a peak at frequency $= 0.047$ $\Omega = 2\pi/(11 \times 12)$ (132 months), which corresponds to a period of $P = 11$ years. This suggests that the number of sunspots follow a periodic cycle with a peak every $P = 11$ years. The general view until

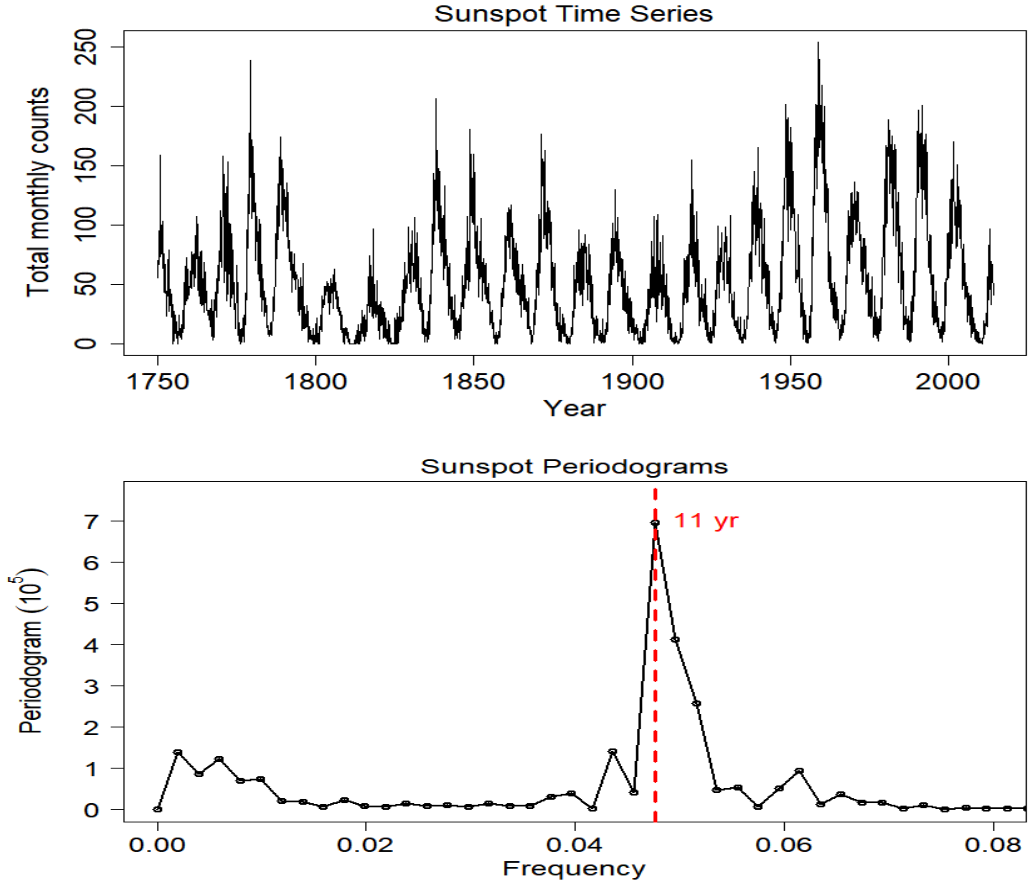


Figure 2.14: Sunspot data from Jan, 1749 to Dec, 2014. There is a peak at about 30 along the line which corresponds to $2\pi/P = 0.047$ and $P \approx 132$ months (11 years).

the 1920s was that most time series were a mix of periodic function with additive noise

$$Y_t = \sum_{j=1}^{P}[A_j \cos(t\Omega_j) + B_j \sin(t\Omega_j)] + \varepsilon_t.$$

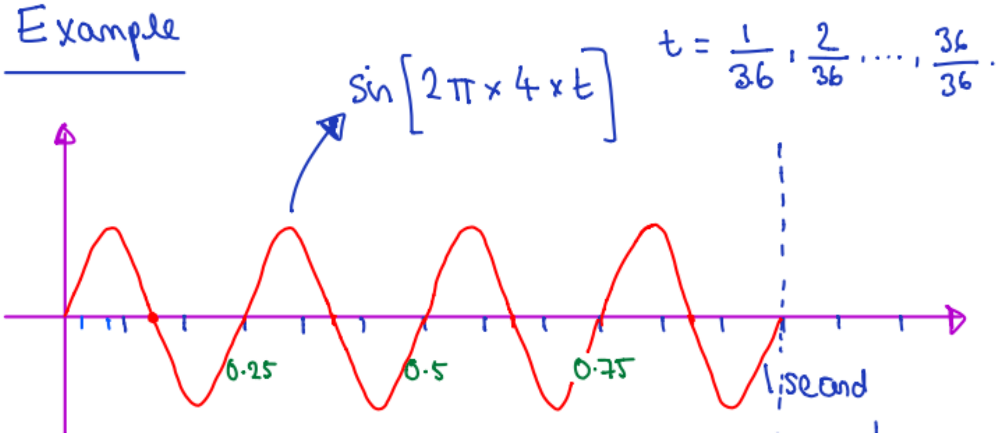
However, in the 1920's, Udny Yule, a statistician, and Gilbert Walker, a Meterologist (working in Pune, India) believed an alternative model could be used to explain the features seen in the periodogram. We consider their proposed approach in Section 4.3.5.

2.6 Data Analysis: EEG data

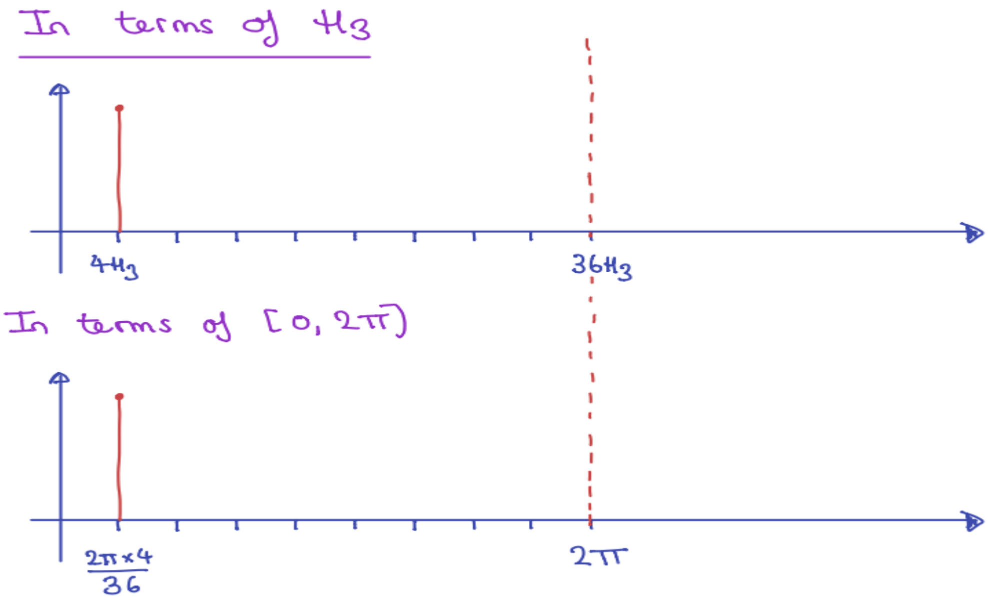
2.6.1 Connecting Hertz and Frequencies

Engineers and neuroscientists often "think" in terms of oscillations or cycles per second. Instead of the sample size they will say the sampling frequency per second (number of observations per second), which is measured in Herz (Hz) and the number of seconds the time series is observed. Thus the periodogram is plotted against cycles per second rather than on the $[0, 2\pi]$ scale. In the following example we connect the two.

Example Suppose that a time series is sampled at 36Hz (36 observations per second) and the signal is $g(u) = \sin(2\pi \times 4u)$ $(u \in \mathbb{R})$. The observed time series in one second is $\{\sin(2\pi \times 4 \times \frac{t}{36})\}_{t=1}^{36}$. An illustration is given below.



We observe from the plot above that period of repetition is $P = 9$ time points (over 36 time points the signal repeats it self every 9 points). Thus in terms of the periodogram this corresponds to a spike at frequency $\omega = 2\pi/9$. But to an engineer this means 4 repetitions a second and a spike at $4Hz$. It is the same plot, just the x-axis is different. The two plots are given below.



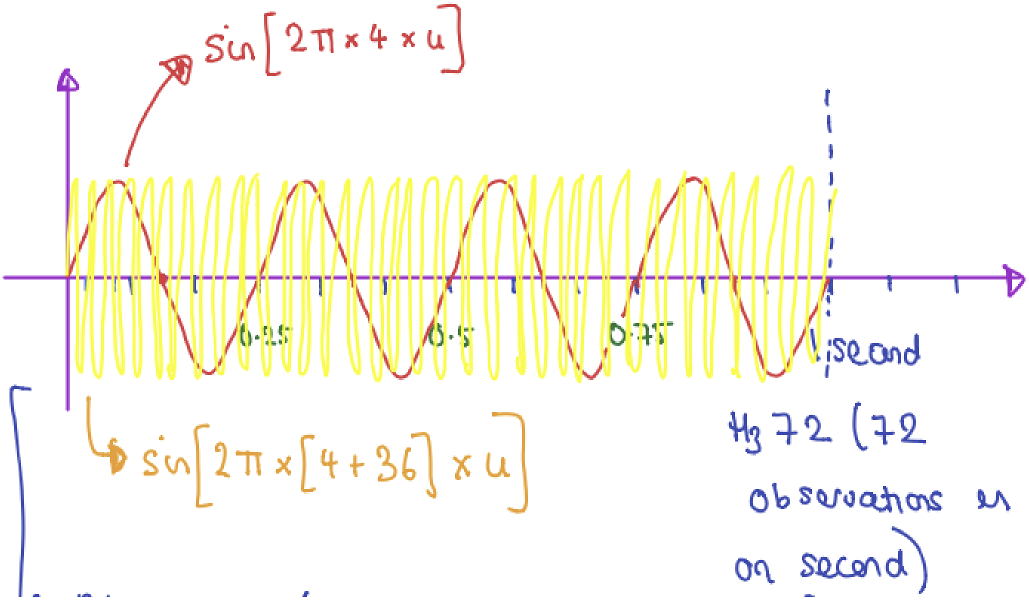
Analysis from the perspective of time series Typically, in time series, the sampling frequency is kept the same. Just the same number of second that the time series is observed grows. This allows us obtain a finer frequency grid on $[0, 2\pi]$ and obtain a better resolution in terms of peaks in frequencies. However, it does not allow is to identify frequencies that are sampled at a higher frequency than the sampling rate.

Returning to the example above. Suppose we observe another signal $h(u) = \sin(2\pi \times (4+36)u)$. If the sampling frequency is 36Hz and $u = 1/36, 2/36, \dots, 36/36$, then

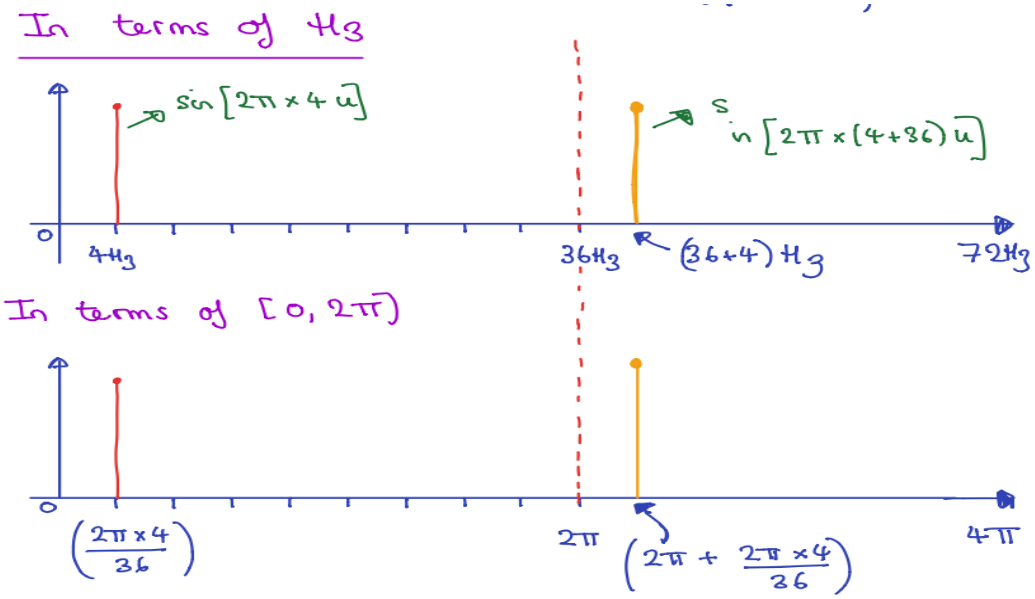
$$\sin\left(2\pi \times 4 \times \frac{t}{36}\right) = \sin\left(2\pi \times (4+36) \times \frac{t}{36}\right) \text{ for all } t \in \mathbb{Z}$$

Thus we cannot tell the differences between these two signals when we sample at 36Hz, even if the observed time series is very long. This is called aliasing.

Analysis from the perspective of an engineer An engineer may be able to improve the hardware and sample the time series at a higher temporal resolution, say, 72Hz. At this higher temporal resolution, the two functions $g(u) = \sin(2\pi \times 4 \times u)$ and $h(u) = \sin(2\pi(4+36)u)$ are different.



In the plot above the red line is $g(u) = \sin(2\pi 4u)$ and the yellow line is $g(u) = \sin(2\pi(4 + 36)u)$. The periodogram for both signals $g(u) = \sin(2\pi \times 4 \times u)$ and $h(u) = \sin(2\pi(4+36)u)$ is given below.



In Hz, we extend the x-axis to include more cycles. The same thing is done for the frequency $[0, 2\pi]$ we extend the frequency range to include higher frequencies. Thus when we observe on a finer temporal grid, we are able to identify higher frequencies. Extending this idea, if we observe time on $\mathbb{R}$, then we can identify all frequencies on $\mathbb{R}$ not just on $[0, 2\pi]$.

2.6.2 Data Analysis

In this section we conduct a preliminary analysis of an EEG data set. A plot of one EEG of one participant at one channel (probe on skull) over 2 seconds (about 512 observations, 256 Hz) is given in Figure 2.15. The neuroscientists who analysis such data use the periodogram to associate the EEG to different types of brain activity. A plot of the periodogam is given Figure 2.16. The periodogram is given in both $[0, \pi]$ and Hz (cycles per second). Observe that the EEG contains a large amount of low frequency information, this is probably due to the slowly changing trend in the original EEG. The neurologists have banded the cycles into bands and associated to each band different types of brain activity (see `https://en.wikipedia.org/wiki/Alpha_wave#Brain_waves`). Very low frequency waves, such as delta, theta and to some extent alpha waves are often associated with low level brain activity (such as breathing). Higher frequencies (alpha and gamma waves) in the EEG are often associated with conscious thought (though none of this is completely understood and there are many debates on this). Studying the periodogram of the EEG in Figures 2.15 and 2.16, we observe that the low frequency information dominates the signal. Therefore, the neuroscientists prefer to decompose the signal into different frequency bands to isolate different parts of the signal. This is usually done by means of a band filter.

As mentioned above, higher frequencies in the EEG are believed to be associated with conscious thought. However, the lower frequencies dominate the EEG. Therefore to put a “microscope” on the higher frequencies in the EEG we isolate them by removing the lower delta and theta band information. This allows us to examine the higher frequencies without being “drowned out” by the more prominent lower frequencies (which have a much larger amplitude). In this data example, we use a Butterworth filter which removes most of the low frequency and very high information (by convolving the original signal with a filter, see Remark 2.6.1). A plot of the periodogam of the orignal EEG together with the EEG after processing with a filter is given in Figure 2.17. Except for a few artifacts (since the Butterworth filter is a finite impulse response filter, and thus only has a finite number of non-zero coefficients), the filter has completely removed the very low frequency information, from $0 - 0.2$ and for the higher frequencies beyond 0.75; we see from the lower plot in Figure

2.17 this means the focus is on 8-32Hz (Hz = number of cycles per second). We observe that most of the frequencies in the interval $[0.2, 0.75]$ have been captured with only a slight amount of distortion. The processed EEG after passing it through the filter is given in Figure 2.18, this data set corresponds to the red periodogram plot seen in Figure 2.17. The corresponding processed EEG clearly shows the evidence of pseudo frequencies described in the section above, and often the aim is to model this processed EEG.

The plot of the original, filtered and the differences in the EEG is given in Figure 2.19. We see the difference (bottom plot) contains the trend in the original EEG and also the small very high frequency fluctuations (probably corresponding to the small spike in the original periodogram in the higher frequencies).

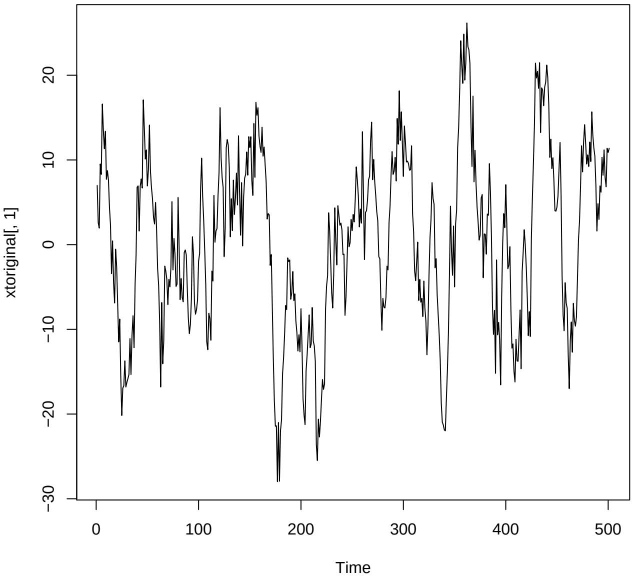


Figure 2.15: Original EEG..

Remark 2.6.1 (How filtering works) *A linear filter is essentially a linear combination of the time series with some weights. The weights are moved along the time series. For example, if $\{h_k\}$ is the filter. Then the filtered time series $\{X_t\}$ is the convolution*

$$Y_t = \sum_{s=0}^{\infty} h_s X_{t-s},$$

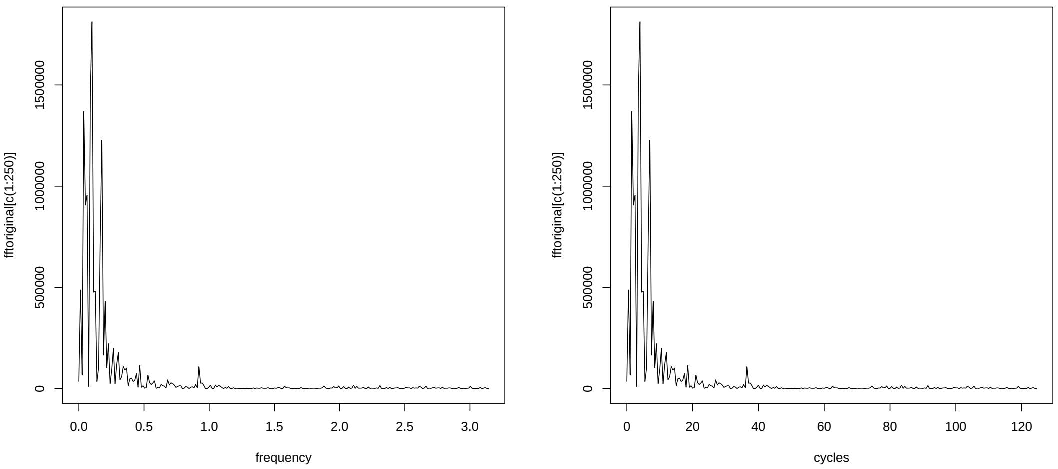


Figure 2.16: Left: Periodogram of original EEG on $[0, 2\pi]$. Right: Periodogram in terms of cycles per second.

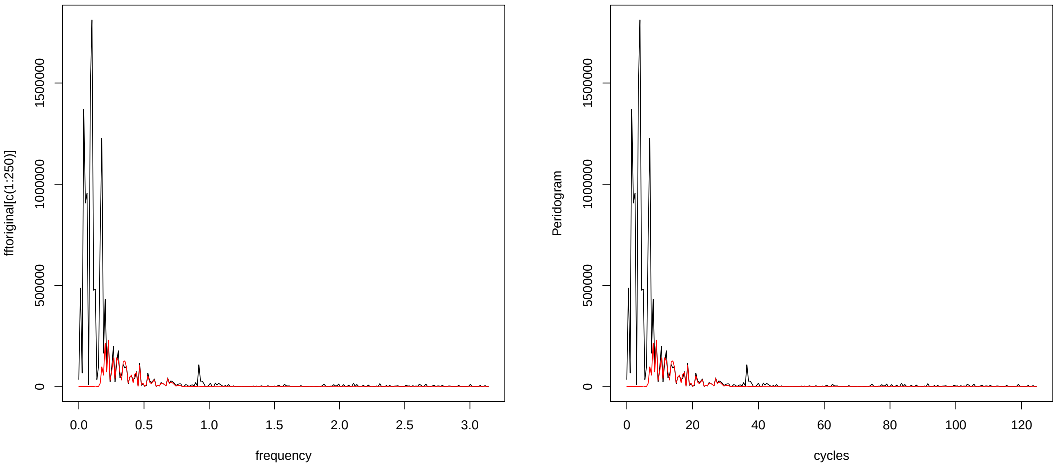


Figure 2.17: The periodogram of original EEG overlayed with processed EEG (in red). The same plot is given below, but the x-axis corresponds to cycles per second (measured in Hz)

note that h_s can be viewed as a moving window. However, the moving window (filter) considered in Section **??** *"smooth" and is used to isolate low frequency trend (mean) behaviour. Whereas the general filtering scheme described above can isolate any type of frequency behaviour. To isolate high frequencies the weights $\{h_s\}$ should not be smooth (should not change slowly over k). To understand the impact $\{h_s\}$ has on $\{X_t\}$ we evaluate the Fourier transform of $\{Y_t\}$.*

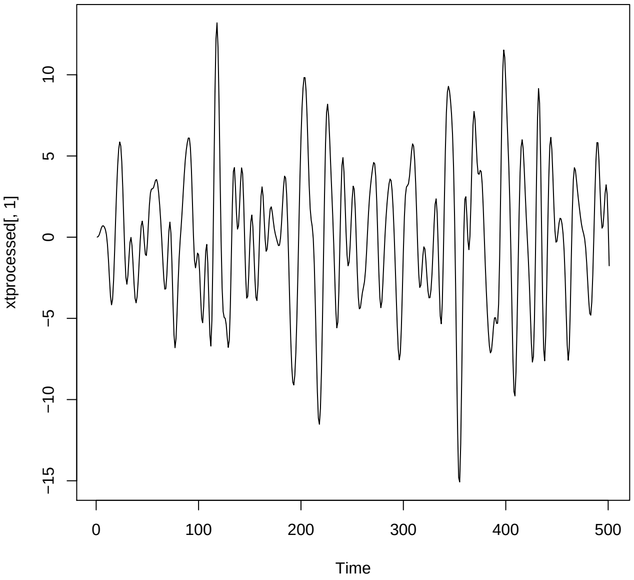


Figure 2.18: Time series after processing with a Buttersworth filter.

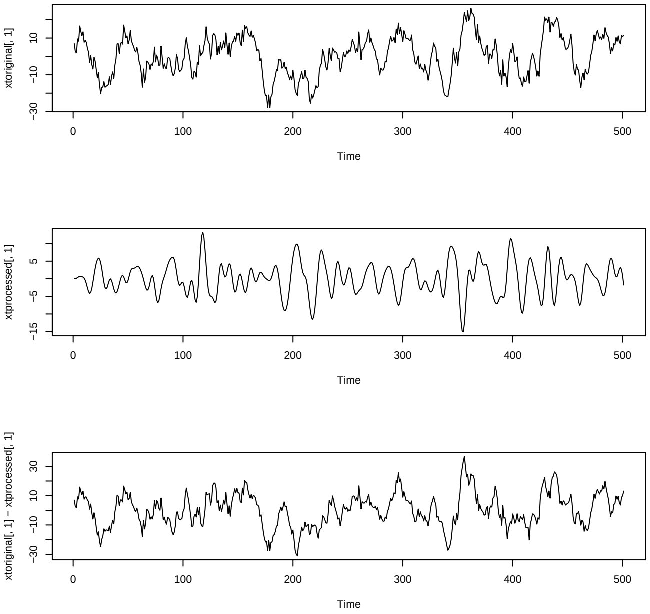


Figure 2.19: Top: Original EEG. Middle: Filtered EEG and Bottom: Difference between Original and Filtered EEG

The periodogram of $\{Y_t\}$ *is*

$$|J_Y(\omega)|^2 = \left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} Y_t e^{it\omega}\right| = \left|\sum_{s=1}^{n} h_s e^{is\omega}\right|^2 \left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t e^{it\omega}\right|^2$$

$$= |H(\omega)|^2 \, |J_X(\omega)|^2 .$$

If $H(\omega)$ is close to zero at certain frequencies it is removing those frequencies in $\{Y_t\}$. Hence using the correct choice of h_s we can isolate certain frequency bands.

Note, if a filter is finite (only a finite number of coefficients), then it is impossible to make the function drop from zero to one. But one can approximately the step by a smooth function (see ***https://en.wikipedia.org/wiki/Butterworth_filter****).*

Remark 2.6.2 *An interesting application of frequency analysis is in the comparison of people in medative and non-medative states (see Gaurav et al. (2019)). A general science video is given in this link.*

2.7 Exercises

Exercise 2.4 (Understanding Fourier transforms) *(i) Let $Y_t = 1$. Plot the Periodogram of $\{Y_t; t = 1, \ldots, 128\}$.*

(ii) Let $Y_t = 1 + \varepsilon_t$, where $\{\varepsilon_t\}$ are iid standard normal random variables. Plot the Periodogram of $\{Y_t; t = 1, \ldots, 128\}$.

(iii) Let $Y_t = \mu(\frac{t}{128})$ where $\mu(u) = 5 \times (2u - 2.5u^2) + 20$. Plot the Periodogram of $\{Y_t; t = 1, \ldots, 128\}$.

(iv) Let $Y_t = 2 \times \sin(\frac{2\pi t}{8})$. Plot the Periodogram of $\{Y_t; t = 1, \ldots, 128\}$.

(v) Let $Y_t = 2 \times \sin(\frac{2\pi t}{8}) + 4 \times \cos(\frac{2\pi t}{12})$. Plot the Periodogram of $\{Y_t; t = 1, \ldots, 128\}$.

You can locate the maximum by using the function `which.max`

Exercise 2.5 *This exercise is aimed at statistics graduate students (or those who have studied STAT613). If you are not a statistics graduate, then you may want help from a statistics student.*

(i) Let

$$\mathcal{S}_n(A, B, \Omega) = \left(\sum_{t=1}^{n} Y_t^2 - 2\sum_{t=1}^{n} Y_t\big(A\cos(\Omega t) + B\sin(\Omega t)\big) + \frac{1}{2}n(A^2 + B^2) \right).$$

Show that

$$2\mathcal{L}_n(A,B,\Omega)+\mathcal{S}_n(A,B,\Omega) = -\frac{(A^2-B^2)}{2}\sum_{t=1}^{n}\cos(2t\Omega) - AB\sum_{t=1}^{n}\sin(2t\Omega).$$

and thus $|\mathcal{L}_n(A,B,\Omega)+\frac{1}{2}\mathcal{S}_n(A,B,\Omega)| = O(1)$ *(ie. the difference does not grow with n).*

Since $\mathcal{L}_n(A,B,\Omega)$ *and* $-\frac{1}{2}\mathcal{S}_n(A,B,\Omega)$ *are asymptotically equivalent (i) shows that we can maximise* $\frac{-1}{2}\mathcal{S}_n(A,B,\Omega)$ *instead of the likelihood* $\mathcal{L}_n(A,B,\Omega)$.

(ii) By profiling out the parameters A and B, use the the profile likelihood to show that $\widehat{\Omega}_n = \arg\max_{\omega}|\sum_{t=1}^{n}Y_t\exp(it\omega)|^2$.

(iii) By using the identity (which is the one-sided Dirichlet kernel)

$$\sum_{t=1}^{n}\exp(i\Omega t) = \begin{cases} \frac{\exp(\frac{1}{2}i(n+1)\Omega)\sin(\frac{1}{2}n\Omega)}{\sin(\frac{1}{2}\Omega)} & 0<\Omega<2\pi \\ n & \Omega = 0 \textit{ or } 2\pi. \end{cases} \tag{2.20}$$

we can show that for $0<\Omega<2\pi$ *we have*

$$\begin{aligned} \sum_{t=1}^{n} t\cos(\Omega t) = O(n) &\quad \sum_{t=1}^{n} t\sin(\Omega t) = O(n) \\ \sum_{t=1}^{n} t^2\cos(\Omega t) = O(n^2) &\quad \sum_{t=1}^{n} t^2\sin(\Omega t) = O(n^2). \end{aligned}$$

Using the above identities, show that the Fisher Information of $\mathcal{L}_n(A,B,\omega)$ *(denoted as* $I(A,B,\omega)$*) is asymptotically equivalent to*

$$2I(A,B,\Omega) = E\big(\frac{\partial^2\mathcal{S}_n}{\partial\omega^2}\big) = \begin{pmatrix} n & 0 & \frac{n^2}{2}B+O(n) \\ 0 & n & -\frac{n^2}{2}A+O(n) \\ \frac{n^2}{2}B+O(n) & -\frac{n^2}{2}A+O(n) & \frac{n^3}{3}(A^2+B^2)+O(n^2) \end{pmatrix}.$$

(iv) Use the Fisher information to show that $|\widehat{\Omega}_n - \Omega| = O(n^{-3/2})$.

Exercise 2.6 *(i) Simulate one hundred times from model* $Y_t = 2\sin(2pit/8)+\varepsilon_t$ *where*

$t = 1, \ldots, n = 60$ and ε_t are iid normal random variables. For each sample, estimate ω, A and B. You can estimate ω, A and B using both nonlinear least squares **and** *also the max periodogram approach described in the previous question.*

For each simulation study obtain the empirical mean squared error $\frac{1}{100}\sum_{i=1}^{100}(\hat{\theta}_i - \theta)^2$ (where θ denotes the parameter and $\hat{\theta}_i$ the estimate).

Note that the more times you simulate the more accurate the empirical standard error will be. The empirical standard error also has an error associated with it, that will be of order $O(1/\sqrt{\text{number of simulations}})$.

Hint 1: When estimating ω restrict the search to $\omega \in [0, \pi]$ (not $[0, 2\pi]$). Also when estimating ω using the max periodogram approach (and A and B) do the search over two grids (a) $\omega = [2\pi j/60, j = 1, \ldots, 30]$ and (b) a finer grid $\omega = [2\pi j/600, j = 1, \ldots, 300]$. Do you see any difference in in your estimates of A, B and Ω over the different grids?

Hint 2: What do you think will happen if the model were changed to $Y_t = 2\sin(2\pi t/10) + \varepsilon_t$ for $t = 1, \ldots, 60$ and the maxim periodogram approach were used to estimate the frequency $\Omega = 2\pi/20$.

(ii) *Repeat the above experiment but this time using the sample size $n = 300$. Compare the quality/MSE of the estimators of A, B and Ω with those in part (i).*

(iii) *Do the same as above (using sample size $n = 60$ and 300) but now use coloured noise given in (2.19) as the errors. How do your estimates compare with (i) and (ii)?*

Hint: A method for simulating dependent data is to use the arima.sim command `ar2 = arima.sim(list(order=c(2,0,0), ar = c(1.5, -0.75)), n=60)`. *This command simulates an $AR(2)$ time series model $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$ (where ε_t are iid normal noise).*

R Code

Simulation and periodogram for model (2.18) with iid errors:

```
temp <- rnorm(128)
signal <- 2*sin(2*pi*c(1:128)/8) + temp # this simulates the series
```

```
# Use the command fft to make the periodogram
P <- abs(fft(signal)/128)**2
frequency <- 2*pi*c(0:127)/128
# To plot the series and periodogram
par(mfrow=c(2,1))
plot.ts(signal)
plot(frequency, P,type="o")
# The estimate of the period is
K1 = which.max(P)
# Phat is the period estimate
Phat = 128/(K1-1)
# To obtain a finer resolution. Pad temp with zeros.
signal2 = c(signal,c(128*9))
frequency2 <- 2*pi*c(0:((128*10)-1))/1280
P2 <- abs(fft(signal2))**2
plot(frequency2, P2 ,type="o")
# To estimate the period we use
K2 = which.max(P)
# Phat2 is the period estimate
Phat2 = 1280/(K2-1)
```

Simulation and periodogram for model (2.18) with correlated errors:

```
set.seed(10)
ar2 <- arima.sim(list(order=c(2,0,0), ar = c(1.5, -0.75)), n=128)
signal2 <- 1.5*sin(2*pi*c(1:128)/8) + ar2
P2 <- abs(fft(signal2)/128)**2
frequency <- 2*pi*c(0:127)/128
par(mfrow=c(2,1))
plot.ts(signal2)
plot(frequency, P2,type="o")
```

Chapter 3

Stationary Time Series

3.1 Preliminaries

The past two chapters focussed on the data. It did not study the properties at the population level (except for a brief discussion on period estimation). By population level, we mean what would happen if the sample size is "infinite". We formally define the tools we will need for such an analysis below.

Different types of convergence

(i) Almost sure convergence: $X_n \overset{\text{a.s.}}{\to} a$ as $n \to \infty$ (in this course a will always be a constant). This means for every $\omega \in \Omega$ $X_n(\omega) \to a$, where $P(\Omega) = 1$ as $n \to \infty$ (this is classical limit of a sequence, see Wiki for a definition).

(ii) Convergence in probability: $X_n \overset{\mathcal{P}}{\to} a$. This means that for every $\varepsilon > 0$, $P(|X_n - a| > \varepsilon) \to 0$ as $n \to \infty$ (see Wiki)

(iii) Convergence in mean square $X_n \overset{2}{\to} a$. This means $\mathrm{E}|X_n - a|^2 \to 0$ as $n \to \infty$ (see Wiki).

(iv) Convergence in distribution. This means the distribution of X_n converges to the distribution of X, ie. for all x where F_X is continuous, we have $F_n(x) \to F_X(x)$ as $n \to \infty$ (where F_n and F_X are the distribution functions of X_n and X respectively). This is the simplest definition (see Wiki).

- Implies:
  - (i), (ii) and (iii) imply (iv).

  - (i) implies (ii).
  - (iii) implies (ii).

- Comments:
  - Central limit theorems require (iv).
  - It is often easy to show (iii) (since this only requires mean and variance calculations).

The "$O_p(\cdot)$" notation.

- We use the notation $|\widehat{\theta}_n - \theta| = O_p(n^{-1/2})$ if there exists a random variable A (which does not depend on n) such that $|\widehat{\theta}_n - \theta| \leq An^{-1/2}$.

  Example of when you can use $O_p(n^{-1/2})$. If $\mathrm{E}[\widehat{\theta}_n] = 0$ but $\mathrm{var}[\widehat{\theta}_n] \leq Cn^{-1}$. Then we can say that $\mathrm{E}|\widehat{\theta} - \theta| \leq Cn^{-1/2}$ and thus $|\widehat{\theta} - \theta| = O_p(n^{-1/2})$.

Definition of expectation

- Suppose X is a random variable with density f_X, then

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

  If $\mathrm{E}[X_i] = \mu$, then the sample mean $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ is an (unbiased) estimator of μ (unbiased because $\mathrm{E}[\bar{X}] = \mu$); most estimators will have a bias (but often it is small).

- Suppose (X, Y) is a bivariate random variable with joint density $f_{X,Y}$, then

$$\mathrm{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x, y) dx dy.$$

Definition of covariance

- The covariance is defined as

$$\mathrm{cov}(X, Y) = \mathrm{E}\left((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

- The variance is $\mathrm{var}(X) = \mathrm{E}(X - \mathrm{E}(X))^2 = \mathrm{E}(X^2) = \mathrm{E}(X)^2$.

- Observe $\mathrm{var}(X) = \mathrm{cov}(X, X)$.

- Rules of covariances. If a, b, c are finite constants and X, Y, Z are random variables with $\mathrm{E}(X^2) < \infty$, $\mathrm{E}(Y^2) < \infty$ and $\mathrm{E}(Z^2) < \infty$ (which immediately implies their means are finite). Then the covariance satisfies the linearity property

$$\mathrm{cov}\,(aX + bY + c, Z) = a\mathrm{cov}(X, Z) + b\mathrm{cov}(Y, Z).$$

  Observe the shift c plays no role in the covariance (since it simply shifts the data).

- The variance of vectors. Suppose that A is a matrix and $\underline{X}$ a random vector with variance/-covariance matrix Σ. Then

$$\mathrm{var}(A\underline{X}) = A\mathrm{var}(\underline{X})A' = A\Sigma A', \tag{3.1}$$

  which can be proved using the linearity property of covariances.

- The correlation between X and Y is

$$\mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}$$

  and lies between $[-1, 1]$. If $\mathrm{var}(X) = \mathrm{var}(Y)$ then $\mathrm{cor}(X, Y)$ is the coefficient of the best linear predictor of X given Y and visa versa.

What is covariance and correlation The covariance and correlation measure the linear dependence between two random variables. If you plot realisations of the bivariate random variable (X, Y) (X on x-axis and Y on y-axis), then the best line of best fit

$$\widehat{Y} = \beta_0 + \beta_1 X$$

gives the best linear predictor of Y given X. β_1 is closely related to the covariance. To see how, consider the following example. Given the observation $\{(X_i, Y_i); i = 1, \ldots, n\}$ the gradient of the linear of the line of best fit is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

As the sample size $n \to \infty$ we recall that

$$\widehat{\beta}_1 \overset{\mathcal{P}}{\to} \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(Y)} = \beta_1.$$

$\beta_1 = 0$ if and only if $\mathrm{cov}(X,Y) = 0$. The covariance between two random variables measures the amount of predictive information (in terms of linear prediction) one variable contains about the other. The coefficients in a regression are not symmetric i.e. $P_X(Y) = \beta_1 X$, whereas $P_Y(X) = \gamma_1 Y$ and in general $\beta_1 \neq \gamma_1$. The correlation

$$\mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}$$

is a symmetric measure of dependence between the two variables.

Exercise 3.1 (Covariance calculations practice) *Suppose $\{\varepsilon_t\}$ are uncorrelated random variables with $\mathrm{E}[\varepsilon_t] = 0$ and $\mathrm{E}[\varepsilon_t^2] = \sigma^2$*

- *Let $X_t = \varepsilon_t + 0.5\varepsilon_{t-1}$. Evaluate $\mathrm{cov}(X_t, X_{t+r})$ for $r = 0, \pm1, \pm2, \pm3, \pm4, \pm5$.*
- *Let $X_t = \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j}$ where $|\rho| < 1$. Evaluate $\mathrm{cov}(X_t, X_{t+r})$ for $r \in \mathbb{Z}$ ($0, \pm1, \pm2, \pm3, \pm4, \ldots$).*

Cumulants: A measure of higher order dependence The covariance has a very simple geometric interpretation. But it only measures linear dependence. In time series and many applications in signal processing, more general measures of dependence are needed. These are called cumulants and can simultaneously measure dependence between several variables or variables with themselves. They generalize the notion of a covariance, but as far as I am aware don't have the nice geometric interpretation that a covariance has.

3.1.1 Formal definition of a time series

When we observe the time series $\{x_t\}$, usually we assume that $\{x_t\}$ is a realisation from a random process $\{X_t\}$. We formalise this notion below. The random process $\{X_t; t \in \mathbb{Z}\}$ (where $\mathbb{Z}$ denotes the integers) is defined on the probability space $\{\Omega, \mathcal{F}, P\}$. We explain what these mean below:

(i) Ω is the set of all possible outcomes. Suppose that $\omega \in \Omega$, then $\{X_t(\omega)\}$ is one realisation from the random process. For any given ω, $\{X_t(\omega)\}$ is not random. In time series we will usually assume that what we observe $x_t = X_t(\omega)$ (for some ω) is a typical realisation. That is, for any other $\omega^* \in \Omega$, $X_t(\omega^*)$ will be different, but its general or overall characteristics will be similar.

(ii) $\mathcal{F}$ is known as a sigma algebra. It is a set of subsets of Ω (though not necessarily the set of all subsets, as this can be too large). But it consists of all sets for which a probability can be assigned. That is if $A \in \mathcal{F}$, then a probability is assigned to the set A.

(iii) P is the probability measure over the sigma-algebra $\mathcal{F}$. For every set $A \in \mathcal{F}$ we can define a probability $P(A)$.

There are strange cases, where there is a subset of Ω, which is not in the sigma-algebra $\mathcal{F}$, where $P(A)$ is not defined (these are called non-measurable sets). In this course, we not have to worry about these cases.

This is a very general definition. But it is too general for modelling. Below we define the notion of stationarity and weak dependence, that allows for estimators to have a meaningful interpretation.

3.2 The sample mean and its standard error

We start with the simplest case, estimating the mean when the data is dependent. This is usually estimated with the sample mean. However, for the sample mean to be estimating something reasonable we require a very weak form of stationarity. That is the time series has the same mean for all t i.e.

$$X_t = \underbrace{\mu}_{=\mathrm{E}(X_t)} + \underbrace{(X_t - \mu)}_{=\varepsilon_t},$$

where $\mu = \mathrm{E}(X_t)$ for all t. This is analogous to say that the independent random variables $\{X_t\}$ all have a common mean. Under this assumption $\bar{X}$ is an unbiased estimator of μ. Next, our aim is to obtain conditions under which $\bar{X}$ is a "reasonable" estimator of the mean.

Based on just one realisation of a time series we want to make inference about the parameters associated with the process $\{X_t\}$, such as the mean. We recall that in classical statistics we usually assume we observe several independent realisations, $\{X_t\}$ all with the same distribution, and use $\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ to estimate the mean. Roughly speaking, with several independent realisations we are able to sample over the entire probability space and thus obtain a "good" (meaning consistent or close to true mean) estimator of the mean. On the other hand, if the samples were highly dependent, then it is likely that $\{X_t\}$ is concentrated over a small part of the probability space. In this case, the sample mean will not converge to the mean (be close to the true mean) as the sample size grows.

The mean squared error a measure of closeness One classical measure of closeness between an estimator and a parameter is the mean squared error

$$\mathrm{E}\left[\widehat{\theta}_n - \theta\right]^2 = \mathrm{var}(\widehat{\theta}_n) + \left[\mathrm{E}(\widehat{\theta}_n) - \theta\right]^2.$$

If the estimator is an unbiased estimator of θ then

$$\mathrm{E}\left[\widehat{\theta}_n - \theta\right]^2 = \mathrm{var}(\widehat{\theta}_n).$$

Returning to the sample mean example suppose that $\{X_t\}$ is a time series wher $\mathrm{E}[X_t] = \mu$ for all t. Then tt is clear that this is an unbiased estimator of μ and

$$\mathrm{E}\left[\bar{X}_n - \mu\right]^2 = \mathrm{var}(\bar{X}_n).$$

To see whether it converges in mean square to μ we evaluate its

$$\mathrm{var}(\bar{X}) = n^{-2}(1,\ldots,1)\underbrace{\mathrm{var}(\underline{X}_n)}_{\text{matrix, }\Sigma}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

where

$$\mathrm{var}(\underline{X}_n) = \begin{pmatrix} \mathrm{cov}(X_1,X_1) & \mathrm{cov}(X_1,X_2) & \mathrm{cov}(X_1,X_3) & \ldots & \mathrm{cov}(X_1,X_n) \\ \mathrm{cov}(X_2,X_1) & \mathrm{cov}(X_2,X_2) & \mathrm{cov}(X_2,X_3) & \ldots & \mathrm{cov}(X_2,X_n) \\ \mathrm{cov}(X_3,X_1) & \mathrm{cov}(X_3,X_2) & \mathrm{cov}(X_3,X_3) & \ldots & \mathrm{cov}(X_3,X_n) \\ \vdots & \vdots & \vdots & \ddots & \ldots \\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_n,X_2) & \ldots & \ldots & \mathrm{cov}(X_n,X_n) \end{pmatrix}.$$

Thus

$$\begin{aligned} \text{var}(\bar{X}) &= \frac{1}{n^2}\sum_{t,\tau=1}^{n}\text{cov}(X_t, X_\tau)\frac{1}{n^2}\sum_{t=1}^{n}\text{var}(X_t) + \frac{2}{n^2}\sum_{t=1}^{n}\sum_{\tau=t+1}^{n}\text{cov}(X_t, X_\tau) \\ &= \frac{1}{n^2}\sum_{t=1}^{n}\text{var}(X_t) + \frac{2}{n^2}\sum_{r=1}^{n-1}\sum_{t=1}^{n-|r|}\text{cov}(X_t, X_{t+r}). \end{aligned} \quad (3.2)$$

A typical time series is a half way house between "fully" dependent data and independent data. Unlike classical statistics, in time series, parameter estimation is based on only <u>one</u> realisation $x_t = X_t(\omega)$ (not multiple, independent, replications). Therefore, it would appear impossible to obtain a good estimator of the mean. However good estimators of the mean are still possible, based on just one realisation of the time series so long as certain assumptions are satisfied (i) the process has a constant mean (a type of stationarity) and (ii) despite the fact that each time series is generated from one realisation there is 'short' memory in the observations. That is, what is observed today, x_t has little influence on observations in the future, x_{t+k} (when k is relatively large). Hence, even though we observe one tragectory, that trajectory traverses much of the probability space. The amount of dependency in the time series determines the 'quality' of the estimator. There are several ways to measure the dependency. We know that the most common is the measure of linear dependency, known as the covariance. Formally, the covariance in the stochastic process $\{X_t\}$ is defined as

$$\text{cov}(X_t, X_{t+k}) = \text{E}\left[(X_t - \text{E}(X_t))\left(X_{t+k} - \text{E}\left(X_{t+k}\right)\right)\right] = \text{E}(X_t X_{t+k}) - \text{E}(X_t)\text{E}(X_{t+k}).$$

Noting that if $\{X_t\}$ has zero mean, then the above reduces to $\text{cov}(X_t, X_{t+k}) = \text{E}(X_t X_{t+k})$.

Remark 3.2.1 (Covariance in a time series) *To illustrate the covariance within a time series setting, we generate the time series*

$$X_t = 1.8\cos\left(\frac{2\pi}{5}\right)X_{t-1} - 0.9^2 X_{t-2} + \varepsilon_t \quad (3.3)$$

for $t = 1, \ldots, n$. A scatter plot of X_t against X_{t+r} for $r = 1, \ldots, 4$ and $n = 200$ is given in Figure 3.1. The corresponding sample autocorrelation (ACF) plot (as defined in equation (3.7) is given in Figure 3.2). Focus on the lags $r = 1, \ldots, 4$ in the ACF plot. Observe that they match what is seen in the scatter plots.

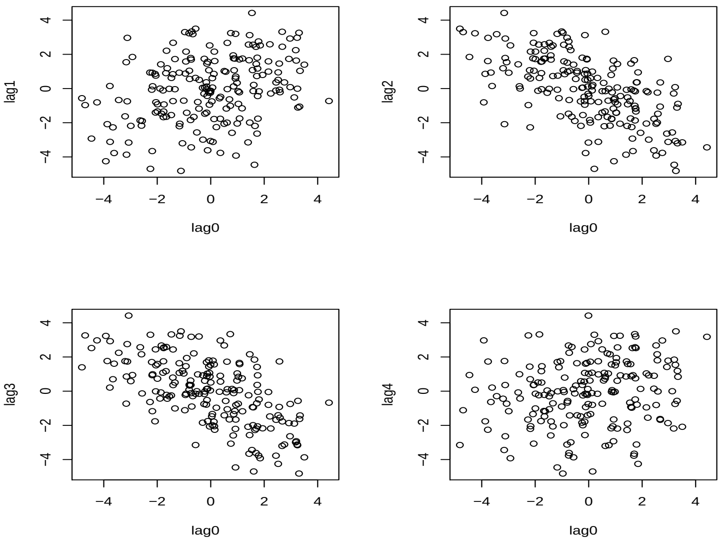


Figure 3.1: From model (3.3). Plot of X_t against X_{t+r} for $r = 1, \dots, 4$. Top left: $r = 1$. Top right: $r = 2$, Bottom left: $r = 3$ and Bottom right: $r = 4$.

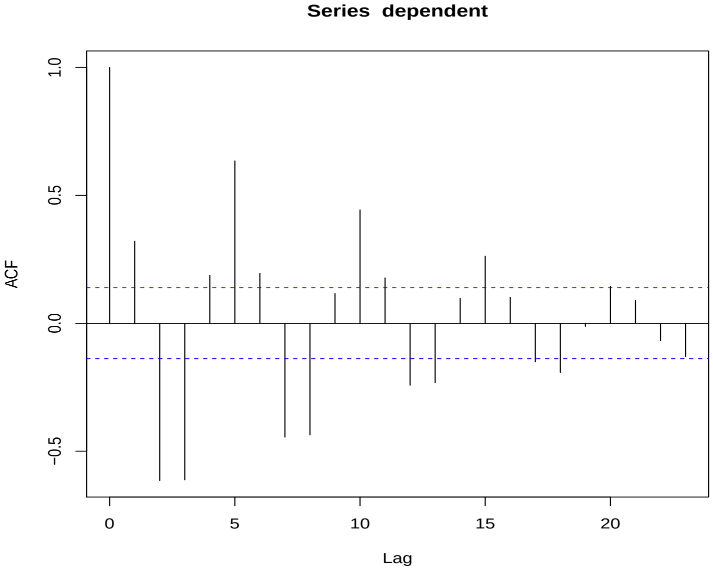


Figure 3.2: ACF plot of realisation from model (3.3).

Using the expression in (3.4) we can deduce under what conditions on the time series we can obtain a reasonable estimator of the mean. If the covariance structure decays at such a rate that the sum of all lags is finite, that is

$$\sup_t \sum_{r=-\infty}^{\infty} |\text{cov}(X_t, X_{t+r})| < \infty,$$

often called short memory), then the variance is

$$
\begin{aligned}
\mathrm{var}(\bar{X}) &\leq \frac{1}{n^2}\sum_{t=1}^{n}\mathrm{var}(X_t) + \frac{2}{n^2}\sum_{r=1}^{n-1}\sum_{t=1}^{n-|r|}|\mathrm{cov}(X_t, X_{t+r})| \\
&\leq \frac{1}{n^2}\sum_{t=1}^{n}\mathrm{var}(X_t) + \frac{2}{n^2}\sum_{t=1}^{n-1}\underbrace{\sum_{r=1}^{\infty}|\mathrm{cov}(X_t, X_{t+r})|}_{\text{finite for all } t \text{ and } n} \leq Cn^{-1} = O(n^{-1}). \qquad (3.4)
\end{aligned}
$$

This rate of convergence is the same as if $\{X_t\}$ were iid/uncorrelated data. However, if the correlations are positive it will be larger than the case that $\{X_t\}$ are uncorrelated.

However, even with this assumption we need to be able to estimate $\mathrm{var}(\bar{X})$ in order to test/-construct CI for μ. Usually this requires the stronger assumption of stationarity, which we define in Section 3.3.

Remark 3.2.2 *It is worth bearing in mind that the covariance only measures linear dependence. For some statistical analysis, such as deriving an expression for the variance of an estimator, the covariance is often sufficient as a measure. However, given* $\mathrm{cov}(X_t, X_{t+k})$ *we cannot say anything about* $\mathrm{cov}(g(X_t), g(X_{t+k}))$*, where g is a nonlinear function. There are occassions where we require a more general measure of dependence (for example, to show asymptotic normality). Examples of more general measures include mixing (and other related notions, such as Mixingales, Near-Epoch dependence, approximate m-dependence, physical dependence, weak dependence), first introduced by Rosenblatt in the 50s (Rosenblatt and Grenander (1997)). In this course we will not cover mixing.*

3.2.1 The variance of the estimated regressors in a linear regression model with correlated errors

Let us return to the parametric models discussed in Section 2.1. The general model is

$$Y_t = \beta_0 + \sum_{j=1}^{p}\beta_j u_{t,j} + \varepsilon_t = \boldsymbol{\beta}'\mathbf{u}_t + \varepsilon_t,$$

where $\mathrm{E}[\varepsilon_t] = 0$ and we will assume that $\{u_{t,j}\}$ are nonrandom regressors. Note this includes the parametric trend models discussed in Section 2.1. We use least squares to estimate $\boldsymbol{\beta}$

$$\mathcal{L}_n(\boldsymbol{\beta}) = \sum_{t=1}^{n}(Y_t - \boldsymbol{\beta}'\mathbf{u}_t)^2,$$

with

$$\hat{\beta}_n = \arg\min \mathcal{L}_n(\boldsymbol{\beta}).$$

Using that

$$\nabla_{\boldsymbol{\beta}}\mathcal{L}_n(\boldsymbol{\beta}) \quad = \quad \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \beta_p} \end{pmatrix} = -2\sum_{t=1}^{n}(Y_t - \boldsymbol{\beta}'\mathbf{u}_t)\mathbf{u}_t,$$

we have

$$\hat{\beta}_n = \arg\min \mathcal{L}_n(\boldsymbol{\beta}) = (\sum_{t=1}^{n}\mathbf{u}_t\mathbf{u}_t')^{-1}\sum_{t=1}^{n}Y_t\mathbf{u}_t,$$

since we solve $\frac{\partial \mathcal{L}_n(\hat{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\beta}} = 0$. To evaluate the variance of $\hat{\boldsymbol{\beta}}_n$ we can either

- Directly evaluate the variance of $\hat{\beta}_n = (\sum_{t=1}^{n}\mathbf{u}_t\mathbf{u}_t')^{-1}\sum_{t=1}^{n}Y_t\mathbf{u}_t$. But this is very special for linear least squares.

- Or use an expansion of $\frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, which is a little longer but generalizes to more complicate estimators and criterions.

We will derive an expression for $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}$. By using $\frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ we can show

$$\begin{aligned} \frac{\partial \mathcal{L}_n(\hat{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\beta}} - \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2\sum_{t=1}^{n}(Y_t - \widehat{\boldsymbol{\beta}}_n'\mathbf{u}_t)\mathbf{u}_t + 2\sum_{t=1}^{n}(Y_t - \boldsymbol{\beta}'\mathbf{u}_t)\mathbf{u}_t \\ &= 2\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right]'\sum_{t=1}^{n}\mathbf{u}_t\mathbf{u}_t'. \end{aligned} \tag{3.5}$$

On the other hand, because $\frac{\partial \mathcal{L}_n(\hat{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\beta}} = 0$ we have

$$\begin{aligned} \frac{\partial \mathcal{L}_n(\hat{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\beta}} - \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -\frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{t=1}^{n}\underbrace{[Y_t - \boldsymbol{\beta}'\mathbf{u}_t]}_{\varepsilon_t}\mathbf{u}_t = \sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t. \end{aligned} \tag{3.6}$$

Equating (3.5) and (3.6) gives

$$\begin{aligned}
\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right]' \sum_{t=1}^{n} \mathbf{u}_t \mathbf{u}_t' &= \sum_{t=1}^{n} \mathbf{u}_t' \varepsilon_t \\
\Rightarrow \left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right] &= \left(\sum_{t=1}^{n} \mathbf{u}_t \mathbf{u}_t'\right)^{-1} \sum_{t=1}^{n} \mathbf{u}_t \varepsilon_t = \left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \mathbf{u}_t'\right)^{-1} \frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \varepsilon_t.
\end{aligned}$$

Using this expression we can see that

$$\mathrm{var}\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right] = \left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \mathbf{u}_t'\right)^{-1} \mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \varepsilon_t\right)\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \mathbf{u}_t'\right)^{-1}.$$

Finally we need only evaluate $\mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \varepsilon_t\right)$ which is

$$\begin{aligned}
\mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \varepsilon_t\right) &= \frac{1}{n^2}\sum_{t,\tau=1}^{n} \mathrm{cov}[\varepsilon_t, \varepsilon_\tau] \mathbf{u}_t \mathbf{u}_\tau' \\
&= \underbrace{\frac{1}{n^2}\sum_{t=1}^{n} \mathrm{var}[\varepsilon_t] \mathbf{u}_t \mathbf{u}_t'}_{\text{expression if independent}} + \underbrace{\frac{1}{n^2}\sum_{t=1}^{n}\sum_{\tau \neq t}^{n} \mathrm{cov}[\varepsilon_t, \varepsilon_\tau] \mathbf{u}_t \mathbf{u}_\tau'}_{\text{additional term due to correlation in the errors}}.
\end{aligned}$$

This expression is analogous to the expression for the variance of the sample mean in (3.4) (make a comparision of the two).

Under the assumption that $\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \mathbf{u}_t'\right)$ is non-singular, $\sup_t \|\mathbf{u}_t\|_1 < \infty$ and $\sup_t \sum_{\tau=-\infty}^{\infty} |\mathrm{cov}(\varepsilon_t, \varepsilon_\tau)| < \infty$, we can see that $\mathrm{var}\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right] = O(n^{-1})$. Estimation of the variance of $\widehat{\boldsymbol{\beta}}_n$ is important and requires one to estimate $\mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \varepsilon_t\right)$. This is often done using the HAC estimator. We describe how this is done in Section 8.5.

3.3 Stationary processes

We have established that one of the main features that distinguish time series analysis from classical methods is that observations taken over time (a time series) can be dependent and this dependency tends to decline the further apart in time these two observations. However, to do any sort of analysis of this time series we have to assume some sort of invariance in the time series, for example the mean or variance of the time series does not change over time. If the marginal distributions of the time series were totally different no sort of inference would be possible (suppose in classical statistics you were given independent random variables all with different distributions, what parameter would you be estimating, it is not possible to estimate anything!).

The typical assumption that is made is that a time series is stationary. Stationarity is a rather intuitive concept, it is an invariant property which means that statistical characteristics of the time series do not change over time. For example, the yearly rainfall may vary year by year, but the average rainfall in two equal length time intervals will be roughly the same as would the number of times the rainfall exceeds a certain threshold. Of course, over long periods of time this assumption may not be so plausible. For example, the climate change that we are currently experiencing is causing changes in the overall weather patterns (we will consider nonstationary time series towards the end of this course). However in many situations, including short time intervals, the assumption of stationarity is quite a plausible. Indeed often the statistical analysis of a time series is done under the assumption that a time series is stationary.

3.3.1 Types of stationarity

There are two definitions of stationarity, weak stationarity which only concerns the covariance of a process and strict stationarity which is a much stronger condition and supposes the distributions are invariant over time.

Definition 3.3.1 (Strict stationarity) *The time series $\{X_t\}$ is said to be strictly stationary if for any finite sequence of integers $t_1, \ldots, t_k$ and shift h the distribution of $(X_{t_1}, \ldots, X_{t_k})$ and $(X_{t_1+h}, \ldots, X_{t_k+h})$ are the same.*

The above assumption is often considered to be rather strong (and given a data it is very hard to check). Often it is possible to work under a weaker assumption called weak/second order stationarity.

Definition 3.3.2 (Second order stationarity/weak stationarity) *The time series $\{X_t\}$ is said to be second order stationary if the mean is constant for all t and if for any t and k the covariance between X_t and X_{t+k} only depends on the lag difference k. In other words there exists a function $c : \mathbb{Z} \to \mathbb{R}$ such that for all t and k we have*

$$c(k) = \text{cov}(X_t, X_{t+k}).$$

Remark 3.3.1 (Strict and second order stationarity) *(i) If a process is strictly stationarity* <u>*and*</u> $\mathrm{E}|X_t^2| < \infty$*, then it is also second order stationary. But the converse is not necessarily true. To show that strict stationarity (with* $\mathrm{E}|X_t^2| < \infty$*) implies second order stationarity, suppose that* $\{X_t\}$ *is a strictly stationary process, then*

$$\begin{aligned}\mathrm{cov}(X_t, X_{t+k}) &= \mathrm{E}(X_t X_{t+k}) - \mathrm{E}(X_t)\mathrm{E}(X_{t+k}) \\ &= \int xy\left[P_{X_t,X_{t+k}}(dx,dy) - P_{X_t}(dx)P_{X_{t+k}}(dy)\right] \\ &= \int xy\left[P_{X_0,X_k}(dx,dy) - P_{X_0}(dx)P_{X_k}(dy)\right] = \mathrm{cov}(X_0, X_k),\end{aligned}$$

where $P_{X_t,X_{t+k}}$ *and* P_{X_t} *is the joint distribution and marginal distribution of* X_t, X_{t+k} *respectively. The above shows that* $\mathrm{cov}(X_t, X_{t+k})$ *does not depend on* t *and* $\{X_t\}$ *is second order stationary.*

(ii) If a process is strictly stationary but the second moment is <u>*not*</u> *finite, then it is not second order stationary.*

(iii) It should be noted that a weakly stationary Gaussian time series is also strictly stationary too (this is the only case where weakly stationary implies strictly stationary).

Example 3.3.1 (The sample mean and its variance under second order stationarity) *Returning the variance of the sample mean discussed (3.4), if a time series is second order stationary, then the sample mean* $\bar{X}$ *is estimating the mean* μ *and the variance of* $\bar{X}$ *is*

$$\begin{aligned}\mathrm{var}(\bar{X}) &= \frac{1}{n^2}\sum_{t=1}^{n}\underbrace{\mathrm{var}(X_t)}_{c(0)} + \frac{2}{n^2}\sum_{r=1}^{n-1}\sum_{t=1}^{n-r}\underbrace{\mathrm{cov}(X_t, X_{t+r})}_{=c(r)} \\ &= \frac{1}{n}c(0) + \frac{2}{n}\sum_{r=1}^{n}\underbrace{\left(\frac{n-r}{n}\right)}_{=1-r/n} c(r),\end{aligned}$$

where we note that above is based on the expansion in (3.4). We approximate the above, by using that the covariances $\sum_r |c(r)| < \infty$*. Therefore for all* r, $(1-r/n)c(r) \to c(r)$ *and* $|\sum_{r=1}^{n}(1-|r|/n)c(r)| \leq \sum_r |c(r)|$*, thus by dominated convergence (see Appendix A)* $\sum_{r=1}^{n}(1-r/n)c(r) \to \sum_{r=1}^{\infty} c(r)$*. This implies that*

$$\mathrm{var}(\bar{X}) \approx \frac{1}{n}c(0) + \frac{2}{n}\sum_{r=1}^{\infty} c(r) = \frac{1}{n}\sum_{r=-\infty}^{\infty} c(r) = O(\frac{1}{n}).$$

The above is often called the long term variance. The above implies that

$$\mathrm{E}(\bar{X}-\mu)^2 = \mathrm{var}(\bar{X}) \to 0, \qquad n \to \infty,$$

which we recall is convergence in mean square. This immediately implies convergence in probability $\bar{X} \xrightarrow{\mathcal{P}} \mu$.

The example above illustrates how second order stationarity gives an elegant expression for the variance and can be used to estimate the standard error associated with $\bar{X}$.

Example 3.3.2 *In Chapter 8 we consider estimation of the autocovariance function. However for now rely on the* `R` *command* `acf`*. For the curious, it evaluates* $\widehat{\rho}(r) = \widehat{c}(r)/\widehat{c}(0)$*, where*

$$\widehat{c}(r) = \frac{1}{n}\sum_{t=1}^{n-r}(X_t - \bar{X})(X_{t+r} - \bar{X}) \tag{3.7}$$

for $r = 1, \ldots, m$ *(*m *is some value that* `R` *defines), you can change the maximum number of lags by using* `acf(data, lag = 30)`*, say). Observe that even if* $X_t = \mu_t$ *(nonconstant mean), from the way* $\widehat{c}(r)$ *(sum of* $(n-r)$ *terms) is defined,* $\widehat{\rho}(r)$ *will decay to zero as* $r \to n$*.*

In Figure 3.3 we give the sample acf plots of the Southern Oscillation Index and the Sunspot data. We observe that are very different. The acf of the SOI decays rapidly, but there does appear to be some sort of 'pattern' in the correlations. On the other hand, there is more "persistence" in the acf of the Sunspot data. The correlations of the acf appear to decay but over a longer period of time and there is a clear periodicity.

Exercise 3.2 *State, with explanation, which of the following time series is second order stationary, which are strictly stationary and which are both.*

(i) $\{\varepsilon_t\}$ *are iid random variables with mean zero and variance one.*

(ii) $\{\varepsilon_t\}$ *are iid random variables from a Cauchy distributon.*

(iii) $X_{t+1} = X_t + \varepsilon_t$*, where* $\{\varepsilon_t\}$ *are iid random variables with mean zero and variance one.*

(iv) $X_t = Y$ *where* Y *is a random variable with mean zero and variance one.*

(iv) $X_t = U_t + U_{t-1} + V_t$*, where* $\{(U_t, V_t)\}$ *is a strictly stationary vector time series with* $\mathrm{E}[U_t^2] < \infty$ *and* $\mathrm{E}[V_t^2] < \infty$*.*

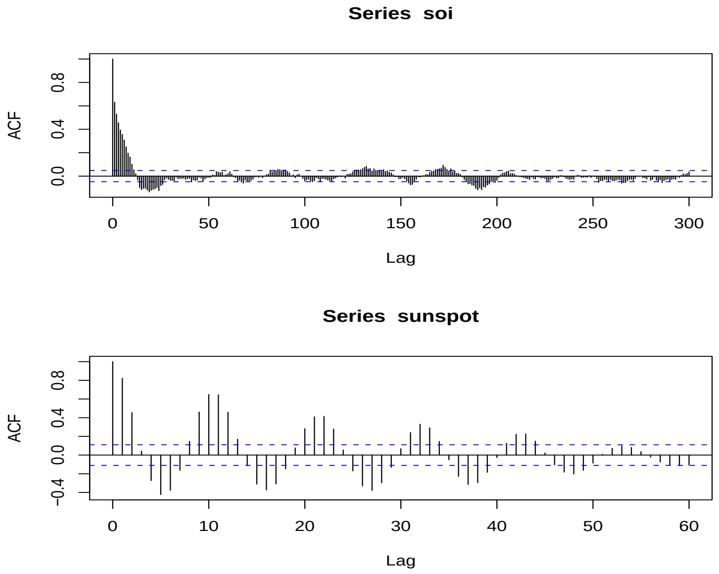


Figure 3.3: Top: ACF of Southern Oscillation data. Bottom ACF plot of Sunspot data.

Exercise 3.3 *(i) Make an ACF plot of the monthly temperature data from 1996-2014.*

(ii) Make and ACF plot of the yearly temperature data from 1880-2013.

(iii) Make and ACF plot of the residuals (after fitting a line through the data (using the command `lsfit(..)$res`*)) of the yearly temperature data from 1880-2013.*

Briefly describe what you see.

Exercise 3.4 *(i) Suppose that $\{X_t\}_t$ is a strictly stationary time series. Let*

$$Y_t = \frac{1}{1+X_t^2}.$$

Show that $\{Y_t\}$ is a second order stationary time series.

(ii) Obtain an approximate expression for the variance of the sample mean of $\{Y_t\}$ in terms of its long run variance (stating the sufficient assumptions for the long run variance to be finite). You do not need to give an analytic expression for the autocovariance, there is not enough information in the question to do this.

(iii) Possibly challenging question. Suppose that

$$Y_t = g(\theta_0, t) + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables and $g(\theta_0, t)$ is a deterministic mean and θ_0 is an unknown parameter. Let

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{t=1}^{n} (Y_t - g(\theta, t))^2.$$

Explain why the quantity

$$\widehat{\theta}_n - \theta_0$$

can be expressed, approximately, as a sample mean. You can use approximations and heuristics here.

Hint: Think derivatives and mean value theorems.

Ergodicity (Advanced)

We now motivate the concept of ergodicity. Conceptionally, this is more difficult to understand than the mean and variance. But it is a very helpful tool when analysing estimators. It allows one to simply replace the sample mean by its expectation without the need to evaluating a variance, which is extremely useful in some situations.

It can be difficult to evaluate the mean and variance of an estimator. Therefore, we may want an alternative form of convergence (instead of the mean squared error). To see whether this is possible we recall that for iid random variables we have the very useful law of large numbers

$$\frac{1}{n} \sum_{t=1}^{n} X_t \overset{\text{a.s.}}{\to} \mu$$

and in general $\frac{1}{n} \sum_{t=1}^{n} g(X_t) \overset{\text{a.s.}}{\to} \mathrm{E}[g(X_0)]$ (if $\mathrm{E}[g(X_0)] < \infty$). Does such a result exists in time series? It does, but we require the slightly stronger condition that a time series is ergodic (which is a slightly stronger condition than the strictly stationary).

Definition 3.3.3 (Ergodicity: Formal definition) *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A transformation $T : \Omega \to \Omega$ is said to be measure preserving if for every set $A \in \mathcal{F}$, $P(T^{-1}A) = P(A)$. Moreover, it is said to be an ergodic transformation if $T^{-1}A = A$ implies that $P(A) = 0$ or 1.*

It is not obvious what this has to do with stochastic processes, but we attempt to make a link. Let us suppose that $X = \{X_t\}$ is a strictly stationary process defined on the probability space $(\Omega, \mathcal{F}, P)$.

By strict stationarity the transformation (shifting a sequence by one)

$$T(x_1, x_2, \ldots) = (x_2, x_3, \ldots),$$

is a measure preserving transformation. To understand ergodicity we define the set A, where

$$A = \{\omega : (X_1(\omega), X_0(\omega), \ldots) \in H\}. = \{\omega : X_{-1}(\omega), \ldots, X_{-2}(\omega), \ldots) \in H\}.$$

The stochastic process is said to be ergodic, if the only sets which satisfies the above are such that $P(A) = 0$ or 1. Roughly, this means there cannot be too many outcomes ω which generate sequences which 'repeat' itself (are periodic in some sense). An equivalent definition is given in (3.8). From this definition is can be seen why "repeats" are a bad idea. If a sequence repeats the time average is unlikey to converge to the mean.

See Billingsley (1994), page 312-314, for examples and a better explanation.

The definition of ergodicity, given above, is quite complex and is rarely used in time series analysis. However, one consequence of ergodicity is the ergodic theorem, which is extremely useful in time series. It states that if $\{X_t\}$ is an ergodic stochastic process then

$$\frac{1}{n}\sum_{t=1}^{n} g(X_t) \overset{\text{a.s.}}{\to} \mathrm{E}[g(X_0)]$$

for any function $g(\cdot)$. And in general for any shift $\tau_1, \ldots, \tau_k$ and function $g : \mathbb{R}^{k+1} \to \mathbb{R}$ we have

$$\frac{1}{n}\sum_{t=1}^{n} g(X_t, X_{t+\tau_1}, \ldots, X_{t+\tau_k}) \overset{\text{a.s.}}{\to} \mathrm{E}[g(X_0, \ldots, X_{t+\tau_k})] \tag{3.8}$$

(often (3.8) is used as the definition of ergodicity, as it is an iff with the ergodic definition). This result generalises the strong law of large numbers (which shows almost sure convergence for iid random variables) to dependent random variables. It is an extremely useful result, as it shows us that "mean-type" estimators consistently estimate their mean (without any real effort). The only drawback is that we do not know the speed of convergence.

(3.8) gives us an idea of what constitutes an ergodic process. Suppose that $\{\varepsilon_t\}$ is an ergodic process (a classical example are iid random variables) then any reasonable (meaning measurable) function of X_t is also ergodic. More precisely, if X_t is defined as

$$X_t = h(\ldots, \varepsilon_t, \varepsilon_{t-1}, \ldots), \tag{3.9}$$

where $\{\varepsilon_t\}$ are iid random variables and $h(\cdot)$ is a measureable function, then $\{X_t\}$ is an Ergodic process. For full details see Stout (1974), Theorem 3.4.5.

Remark 3.3.2 *As mentioned above all Ergodic processes are stationary, but a stationary process is not necessarily ergodic. Here is one simple example. Suppose that $\{\varepsilon_t\}$ are iid random variables and Z is a Bernoulli random variable with outcomes $\{1,2\}$ (where the chance of either outcome is half). Suppose that Z stays the same for all t. Define*

$$X_t = \begin{cases} \mu_1 + \varepsilon_t & Z = 1 \\ \mu_2 + \varepsilon_t & Z = 2. \end{cases}$$

It is clear that $\mathrm{E}(X_t|Z = i) = \mu_i$ and $\mathrm{E}(X_t) = \frac{1}{2}(\mu_1 + \mu_2)$. This sequence is stationary. However, we observe that $\frac{1}{T}\sum_{t=1}^{T} X_t$ will only converge to one of the means, hence we do not have almost sure convergence (or convergence in probability) to $\frac{1}{2}(\mu_1 + \mu_2)$.

`R` code

To make the above plots we use the commands

```
par(mfrow=c(2,1))
acf(soi,lag.max=300)
acf(sunspot,lag.max=60)
```

3.3.2 Towards statistical inference for time series

Returning to the sample mean Example 3.3.1. Suppose we want to construct CIs or apply statistical tests on the mean. This requires us to estimate the long run variance (assuming stationarity)

$$\mathrm{var}(\bar{X}) \approx \frac{1}{n}c(0) + \frac{2}{n}\sum_{r=1}^{\infty} c(r).$$

There are several ways this can be done, either by fitting a model to the data and from the model estimate the covariance or doing it nonparametrically. This example motivates the contents of the course:

(i) Modelling, finding suitable time series models to fit to the data.

(ii) Forecasting, this is essentially predicting the future given current and past observations.

(iii) Estimation of the parameters in the time series model.

(iv) The spectral density function and frequency domain approaches, sometimes within the frequency domain time series methods become extremely elegant.

(v) Analysis of nonstationary time series.

(vi) Analysis of nonlinear time series.

(vii) How to derive sampling properties.

3.4 What makes a covariance a covariance?

The covariance of a stationary process has several very interesting properties. The most important is that it is positive semi-definite, which we define below.

Definition 3.4.1 (Positive semi-definite sequence) *(i) A sequence $\{c(k); k \in \mathbb{Z}\}$ ($\mathbb{Z}$ is the set of all integers) is said to be positive semi-definite if for any $n \in \mathbb{Z}$ and sequence $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ the following is satisfied*

$$\sum_{i,j=1}^{n} c(i-j)x_i x_j \geq 0.$$

(ii) A function is said to be an <u>even</u> positive semi-definite sequence if (i) is satisfied and $c(k) = c(-k)$ for all $k \in \mathbb{Z}$.

An extension of this notion is the positive semi-definite function.

Definition 3.4.2 (Positive semi-definite function) *(i) A function $\{c(u); u \in \mathbb{R}\}$ is said to be positive semi-definite if for any $n \in \mathbb{Z}$ and sequence $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ the following is satisfied*

$$\sum_{i,j=1}^{n} c(u_i - u_j)x_i x_j \geq 0.$$

(ii) A function is said to be an <u>even</u> positive semi-definite function if (i) is satisfied and $c(u) = c(-u)$ for all $u \in \mathbb{R}$.

Remark 3.4.1 *You have probably encountered this positive definite notion before, when dealing with positive definite matrices. Recall the $n \times n$ matrix Σ_n is positive semi-definite if for all $\underline{x} \in \mathbb{R}^n$ $\underline{x}'\Sigma_n\underline{x} \geq 0$. To see how this is related to positive semi-definite matrices, suppose that the matrix Σ_n has a special form, that is the elements of Σ_n are $(\Sigma_n)_{i,j} = c(i-j)$. Then $\underline{x}'\Sigma_n\underline{x} = \sum_{i,j}^n c(i-j)x_ix_j$. We observe that in the case that $\{X_t\}$ is a stationary process with covariance $c(k)$, the variance covariance matrix of $\underline{X}_n = (X_1, \ldots, X_n)$ is Σ_n, where $(\Sigma_n)_{i,j} = c(i-j)$.*

We now take the above remark further and show that the covariance of a stationary process is positive semi-definite.

Theorem 3.4.1 *Suppose that $\{X_t\}$ is a discrete time/continuous stationary time series with covariance function $\{c(k)\}$, then $\{c(k)\}$ is an even positive semi-definite sequence/function. Conversely for any <u>even</u> positive semi-definite sequence/function there exists a stationary time series with this positive semi-definite sequence/function as its covariance function.*

PROOF. We prove the result in the case that $\{X_t\}$ is a discrete time time series, ie. $\{X_t; t \in \mathbb{Z}\}$.

We first show that $\{c(k)\}$ is a positive semi-definite sequence. Consider any sequence $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, and the double sum $\sum_{i,j}^n x_ic(i-j)x_j$. Define the random variable $Y = \sum_{i=1}^n x_iX_i$. It is straightforward to see that $\text{var}(Y) = \underline{x}'\text{var}(\underline{X}_n)\underline{x} = \sum_{i,j=1}^n c(i-j)x_ix_j$ where $\underline{X}_n = (X_1, \ldots, X_n)$. Since for any random variable Y, $\text{var}(Y) \geq 0$, this means that $\sum_{i,j=1}^n x_ic(i-j)x_j \geq 0$, hence $\{c(k)\}$ is a positive definite sequence.

To show the converse, that is for any positive semi-definite sequence $\{c(k)\}$ we can find a corresponding stationary time series with the covariance $\{c(k)\}$ is relatively straightfoward, but depends on defining the characteristic function of a process and using Komologorov's extension theorem. We omit the details but refer an interested reader to Brockwell and Davis (1998), Section 1.5. □

In time series analysis usually the data is analysed by fitting a *model* to the data. The model (so long as it is correctly specified, we will see what this means in later chapters) guarantees the covariance function corresponding to the model (again we cover this in later chapters) is positive definite. This means, in general we do not have to worry about positive definiteness of the covariance function, as it is implicitly implied.

On the other hand, in spatial statistics, often the object of interest is the covariance function and specific classes of covariance functions are fitted to the data. In which case it is necessary to ensure that the covariance function is semi-positive definite (noting that once a covariance function has been found by Theorem 3.4.1 there must exist a spatial process which has this covariance function). It is impossible to check for positive definiteness using Definitions 3.4.1 or 3.4.1. Instead an alternative but equivalent criterion is used. The general result, which does not impose any conditions on $\{c(k)\}$ is stated in terms of positive measures (this result is often called Bochner's theorem). Instead, we place some conditions on $\{c(k)\}$, and state a simpler version of the theorem.

Theorem 3.4.2 *Suppose the coefficients $\{c(k); k \in \mathbb{Z}\}$ are absolutely summable (that is $\sum_k |c(k)| < \infty$). Then the sequence $\{c(k)\}$ is positive semi-definite if an only if the function $f(\omega)$, where*

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} c(k) \exp(ik\omega),$$

is nonnegative for all $\omega \in [0, 2\pi]$.

We also state a variant of this result for positive semi-definite functions. Suppose the function $\{c(u); k \in \mathbb{R}\}$ is absolutely summable (that is $\int_{\mathbb{R}} |c(u)|du < \infty$). Then the function $\{c(u)\}$ is positive semi-definite if and only if the function $f(\omega)$, where

$$f(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} c(u) \exp(iu\omega) du \geq 0$$

for all $\omega \in \mathbb{R}$.

The generalisation of the above result to dimension d is that $\{c(\boldsymbol{u}); \boldsymbol{u} \in \mathbb{R}^d\}$ is a positive semi-definite sequence if and if

$$f(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} c(\boldsymbol{u}) \exp(i\boldsymbol{u}'\boldsymbol{\omega}) d\boldsymbol{u} \geq 0$$

for all $\boldsymbol{\omega}^d \in \mathbb{R}^d$.

PROOF. See Section 10.4.1.

Example 3.4.1 *We will show that sequence $c(0) = 1$, $c(1) = 0.5$, $c(-1) = 0.5$ and $c(k) = 0$ for $|k| > 1$ a positive definite sequence.*

From the definition of spectral density given above we see that the 'spectral density' corresponding *to the above sequence is*

$$f(\omega) = 1 + 2 \times 0.5 \times cos(\omega).$$

Since $|\cos(\omega)| \leq 1$, $f(\omega) \geq 0$, *thus the sequence is positive definite. An alternative method is to find a model which has this as the covariance structure. Let* $X_t = \varepsilon_t + \varepsilon_{t-1}$, *where* ε_t *are iid random variables with* $\mathrm{E}[\varepsilon_t] = 0$ *and* $\mathrm{var}(\varepsilon_t) = 0.5$. *This model has this covariance structure.*

3.5 Spatial covariances (advanced)

Theorem 3.4.2 is extremely useful in finding valid spatial covariances. We recall that $c_d : \mathbb{R}^d \to \mathbb{R}$ is a positive semi-definite covariance (on the spatial plane $\mathbb{R}^d$) if there exists a positive function f_d where

$$c_d(\boldsymbol{u}) = \int_{\mathbb{R}^d} f_d(\boldsymbol{\omega}) \exp(-i\boldsymbol{u}'\boldsymbol{\omega}) d\boldsymbol{\omega} \tag{3.10}$$

for all $\boldsymbol{u} \in \mathbb{R}^d$ (the inverse Fourier transform of what was written). This result allows one to find parametric covariance spatial processes.

However, beyond dimension $d = 1$ (which can be considered a "time series"), there exists conditions stronger than spatial (second order) stationarity. Probably the the most popular is spatial isotropy, which is even stronger than stationarity. A covariance c_d is called spatially isotropic if it is stationary and there exist a function $c : \mathbb{R} \to \mathbb{R}$ such that $c_d(\boldsymbol{u}) = c(\|\boldsymbol{u}\|_2)$. It is clear that in the case $d = 1$, a stationary covariance is isotropic since $\mathrm{cov}(X_t, X_{t+1}) = c(1) = c(-1) == \mathrm{cov}(X_t, X_{t-1}) = \mathrm{cov}(X_{t-1}, X_t)$. For $d > 1$, isotropy is a stronger condition than stationarity. The appeal of an isotropic covariance is that the actual directional difference between two observations *does not* impact the covariance, it is simply the Euclidean distance between the two locations (see picture on board). To show that the covariance $c(\cdot)$ is a valid isotropic covariance in dimension d (that is there exists a positive semi-definite function $c_d : \mathbb{R}^d \to \mathbb{R}$ such that $c(\|\boldsymbol{u}\|) = c_d(\boldsymbol{u})$), conditions analogous but not the same as (3.10) are required. We state them now.

Theorem 3.5.1 *If a covariance* $c_d(\cdot)$ *is isotropic, its corresponding spectral density function* f_d *is also isotropic. That is, there exists a positive function* $f : \mathbb{R} \to \mathbb{R}^+$ *such that* $f_d(\boldsymbol{\omega}) = f(\|\boldsymbol{\omega}\|_2)$.

A covariance $c(\cdot)$ *is a valid isotropic covariance in* $\mathbb{R}^d$ *iff there exists a positive function* $f(\cdot; d)$

defined in $\mathbb{R}^+$ such that

$$c(r) = (2\pi)^{d/2}\int_0^\infty \rho^{d/2}J_{(d/2)-1}(\rho)f(\rho;d)d\rho \tag{3.11}$$

where J_n is the order n Bessel function of the first kind.

PROOF. To give us some idea of where this result came from, we assume the first statement is true and prove the second statement for the case the dimension $d = 2$.

By the spectral representation theorem we know that if $c(u_1, u_r)$ is a valid covariance then there exists a positive function f_2 such that

$$c(u_1, u_2) = \int_{\mathbb{R}^2} f_2(\omega_1, \omega_2)\exp(i\omega_1 u_1 + i\omega_2 u_2)d\omega_1 d\omega_2.$$

Next we change variables moving from Euclidean coordinates to polar coordinates (see `https://en.wikipedia.org/wiki/Polar_coordinate_system`), where $s = \sqrt{\omega_1^2 + \omega_2^2}$ and $\theta = tan^{-1}\omega_1/\omega_2$. In this way the spectral density can be written in terms of $f_2(\omega_1, \omega_2) = f_{P,2}(r, \theta)$ and we have

$$c(u_1, u_2) = \int_0^\infty \int_0^{2\pi} r f_{P,2}(s,\theta)\exp(isu_1\cos\theta + isu_2\sin\theta)dsd\theta.$$

We convert the covariance in terms of polar coordinates $c(u_1, u_2) = c_{P,2}(r, \Omega)$ (where $u_1 = r\cos\Omega$ and $u_2 = r\sin\Omega$) to give

$$\begin{aligned} c_{P,2}(r,\Omega) &= \int_0^\infty \int_0^{2\pi} s f_{P,2}(s,\theta)\exp\left[isr\left(\cos\Omega\cos\theta + \sin\Omega\sin\theta\right)\right]dsd\theta \\ &= \int_0^\infty \int_0^{2\pi} s f_{P,2}(s,\theta)\exp\left[isr\cos\left(\Omega - \theta\Omega\right)\right]dsd\theta. \end{aligned} \tag{3.12}$$

So far we have not used isotropy of the covariance, we have simply rewritten the spectral representation in terms of polar coordinates.

Now, we consider the special case that the covariance is isotropic, this means that there exists a function c such that $c_{P,2}(r, \Omega) = c(r)$ for all r and Ω. Furthermore, by the first statement of the theorem, if the covariance is isotropic, then there exists a positive function $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$f_{P,2}(s,\theta) = f(s)$ for all s and θ. Using these two facts and substituting them into (3.12) gives

$$\begin{aligned} c(r) &= \int_0^\infty \int_0^{2\pi} s f(s) \exp\left[isr\cos\left(\Omega - \theta\Omega\right)\right] ds d\theta \\ &= \int_0^\infty s f(s) \underbrace{\int_0^{2\pi} \exp\left[isr\cos\left(\Omega - \theta\Omega\right)\right] d\theta}_{=2\pi J_0(s)} ds. \end{aligned}$$

For the case, $d = 2$ we have obtained the desired result. Note that the Bessel function $J_0(\cdot)$ is effectively playing the same role as the exponential function in the general spectral representation theorem. □

The above result is extremely useful. It allows one to construct a valid isotropic covariance function in dimension d with a positive function f. Furthermore, it shows that an isotropic covariance $c(r)$ may be valid in dimension in $d = 1, \ldots, 3$, but for $d > 3$ it may not be valid. That is for $d > 3$, there does not exist a positive function $f(\cdot; d)$ which satisfies (3.11). Schoenberg showed that an isotropic covariance $c(r)$ was valid in all dimensions d iff there exists a representation

$$c(r) = \int_0^\infty \exp(-r^2t^2) dF(t),$$

where F is a probability measure. In most situations the above can be written as

$$c(r) = \int_0^\infty \exp(-r^2t^2) f(t) dt,$$

where $f : \mathbb{R}^+ \to \mathbb{R}^+$. This representation turns out to be a very fruitful method for generating parametric families of isotropic covariances which are valid on all dimensions d. These include the Matern class, Cauchy class, Powered exponential family. The feature in common to all these isotropic covariance functions is that all the covariances are strictly positive and strictly decreasing. In other words, the cost for an isotropic covariance to be valid in all dimensions is that it can only model positive, monotonic correlations. The use of such covariances have become very popular in modelling Gaussian processes for problems in machine learning (see `http://www.gaussianprocess.org/gpml/chapters/RW1.pdf`).

For an excellent review see **?**, Section 2.5.

3.6 Exercises

Exercise 3.5 *Which of these sequences can used as the autocovariance function of a second order stationary time series?*

(i) $c(-1) = 1/2$, $c(0) = 1$, $c(1) = 1/2$ *and for all* $|k| > 1$, $c(k) = 0$.

(ii) $c(-1) = -1/2$, $c(0) = 1$, $c(1) = 1/2$ *and for all* $|k| > 1$, $c(k) = 0$.

(iii) $c(-2) = -0.8$, $c(-1) = 0.5$, $c(0) = 1$, $c(1) = 0.5$ *and* $c(2) = -0.8$ *and for all* $|k| > 2$, $c(k) = 0$.

Exercise 3.6 *(i) Show that the function* $c(u) = \exp(-a|u|)$ *where* $a > 0$ *is a positive semi-definite function.*

(ii) Show that the commonly used exponential spatial covariance defined on $\mathbb{R}^2$, $c(u_1, u_2) = \exp(-a\sqrt{u_1^2 + u_2^2})$, *where* $a > 0$, *is a positive semi-definite function.*

Hint: One method is to make a change of variables using Polar coordinates. You may also want to harness the power of Mathematica or other such tools.

Chapter 4

Linear time series

Prerequisites

- Familarity with linear models in regression.
- Find the polynomial equations. If the solution is complex writing complex solutions in polar form $x + iy = re^{i\theta}$, where θ is the phased and r the modulus or magnitude.

Objectives

- Understand what causal and invertible is.
- Know what an AR, MA and ARMA time series model is.
- Know how to find a solution of an ARMA time series, and understand why this is important (how the roots determine causality and why this is important to know - in terms of characteristics in the process and also simulations).
- Understand how the roots of the AR can determine 'features' in the time series and covariance structure (such as pseudo periodicities).

4.1 Motivation

The objective of this chapter is to introduce the linear time series model. Linear time series models are designed to model the covariance structure in the time series. There are two popular subgroups of linear time models (a) the autoregressive and (a) the moving average models, which can be combined to make the autoregressive moving average models.

We motivate the autoregressive from the perspective of classical linear regression. We recall one objective in linear regression is to predict the response variable given variables that are observed. To do this, typically linear dependence between response and variable is assumed and we model Y_i as

$$Y_i = \sum_{j=1}^{p} a_j X_{ij} + \varepsilon_i,$$

where ε_i is such that $\mathrm{E}[\varepsilon_i | X_{ij}] = 0$ and more commonly ε_i and X_{ij} are independent. In linear regression once the model has been defined, we can immediately find estimators of the parameters, do model selection etc.

Returning to time series, one major objective is to predict/forecast the future given current and past observations (just as in linear regression our aim is to predict the response given the observed variables). At least formally, it seems reasonable to represent this as

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t, \qquad t \in \mathbb{Z} \tag{4.1}$$

where we assume that $\{\varepsilon_t\}$ are independent, identically distributed, zero mean random variables. Model (4.1) is called an autoregressive model of order p (AR(p) for short). Further, it would appear that

$$\mathrm{E}(X_t | X_{t-1}, \ldots, X_{t-p}) = \sum_{j=1}^{p} \phi_j X_{t-j}. \tag{4.2}$$

I.e. the expected value of X_t given that $X_{t-1}, \ldots, X_{t-p}$ have already been observed), thus the past values of X_t have a linear influence on the conditional mean of X_t. However (4.2) not necessarily true.

Unlike the linear regression model, (4.1) is an infinite set of linear difference equations. This means, for this systems of equations to be well defined, it needs to have a solution which is meaningful. To understand why, recall that (4.1) is defined for all $t \in \mathbb{Z}$, so let us start the equation at the beginning of time ($t = -\infty$) and run it on. Without any constraint on the parameters $\{\phi_j\}$, there is no reason to believe the solution is finite (contrast this with linear regression where these issues are not relevant). Therefore, the first thing to understand is under what conditions will the AR model (4.1) have a well defined stationary solution and what features in a time series is the solution able to capture.

Of course, one could ask why go through to the effort. One could simply use least squares to estimate the parameters. This is possible, but there are two related problems (a) without a proper analysis it is not clear whether model has a meaningful solution (for example in Section 6.4 we show that the least squares estimator can lead to misspecified models), it's not even possible to make simulations of the process (b) it is possible that $\mathrm{E}(\varepsilon_t|X_{t-p}) \neq 0$, this means that least squares is not estimating ϕ_j and is instead estimating an entirely different set of parameters! Therefore, there is a practical motivation behind our theoretical treatment.

In this chapter we will be deriving conditions for a strictly stationary solution of (4.1). Under these moment conditions we obtain a strictly stationary solution of (4.1). In Chapter 6 we obtain conditions for (4.1) to have both a strictly stationary and second order stationary solution. It is worth mentioning that it is possible to obtain a strictly stationary solution to (4.1) under weaker conditions (see Theorem 13.0.1).

How would you simulate from the following model? One simple method for understanding a model is to understand how you would simulate from it:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-1} + \varepsilon_t \qquad t = \ldots, -1, 0, 1, \ldots.$$

4.2 Linear time series and moving average models

4.2.1 Infinite sums of random variables

Before defining a linear time series, we define the MA(q) model which is a subclass of linear time series. Let us supppose that $\{\varepsilon_t\}$ are iid random variables with mean zero and finite variance. The time series $\{X_t\}$ is said to have a MA(q) representation if it satisfies

$$X_t = \sum_{j=0}^{q} \psi_j \varepsilon_{t-j},$$

where $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{var}(\varepsilon_t) = 1$. It is clear that X_t is a rolling finite weighted sum of $\{\varepsilon_t\}$, therefore $\{X_t\}$ must be well defined. We extend this notion and consider infinite sums of random variables.

Now, things become more complicated, since care must be always be taken with anything involving *infinite sums.* More precisely, for the sum

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

to be well defined (has a finite limit), the partial sums $S_n = \sum_{j=-n}^{n} \psi_j \varepsilon_{t-j}$ should be (almost surely) finite and the sequence S_n should converge (ie. $|S_{n_1} - S_{n_2}| \to 0$ as $n_1, n_2 \to \infty$). A random variable makes no sense if it is infinite. Therefore we must be sure that X_t is finite (this is what we mean by being well defined).

Below, we give conditions under which this is true.

Lemma 4.2.1 *Suppose* $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ *and* $\{X_t\}$ *is a strictly stationary time series with* $\mathrm{E}|X_t| < \infty$. *Then* $\{Y_t\}$*, defined by*

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j},$$

is a strictly stationary time series. Furthermore, the partial sum converges almost surely, $Y_{n,t} = \sum_{j=-n}^{n} \psi_j X_{t-j} \to Y_t$. *If* $\mathrm{var}(X_t) < \infty$*, then* $\{Y_t\}$ *is second order stationary and converges in mean square (that is* $\mathrm{E}(Y_{n,t} - Y_t)^2 \to 0$*).*

PROOF. See Brockwell and Davis (1998), Proposition 3.1.1 or Fuller (1995), Theorem 2.1.1 (page 31) (also Shumway and Stoffer (2006), page 86). □

Example 4.2.1 *Suppose* $\{X_t\}$ *is a strictly stationary time series with* $\mathrm{var}(X_t) < \infty$. *Define* $\{Y_t\}$ *as the following infinite sum*

$$Y_t = \sum_{j=0}^{\infty} j^k \rho^j |X_{t-j}|$$

where $|\rho| < 1$. *Then* $\{Y_t\}$ *is also a strictly stationary time series with a finite variance.*

We will use this example later in the course.

Having derived conditions under which infinite sums are well defined, we can now define the general class of linear and MA(∞) processes.

Definition 4.2.1 (The linear process and moving average (MA)(∞)) *Suppose that* $\{\varepsilon_t\}$ *are* *iid random variables,* $\sum_{j=0}^{\infty}|\psi_j| < \infty$ *and* $\mathrm{E}(|\varepsilon_t|) < \infty$*.*

(i) A time series is said to be a linear time series if it can be represented as

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

where $\{\varepsilon_t\}$ *are iid random variables with finite variance. Note that since that as these sums are well defined by equation (3.9)* $\{X_t\}$ *is a strictly stationary (ergodic) time series.*

This is a rather strong definition of a linear process. A more general definition is $\{X_t\}$ *has the representation*

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

where $\{\varepsilon_t\}$ *are* uncorrelated *random variables with mean zero and variance one (thus the independence assumption has been dropped).*

(ii) The time series $\{X_t\}$ *has a MA*(∞) *representation if it satisfies*

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}. \tag{4.3}$$

[1]

The difference between an MA(∞) process and a linear process is quite subtle. A linear process involves both past, present and future innovations $\{\varepsilon_t\}$, whereas the MA(∞) uses only past and present innovations.

A very interesting class of models which have MA(∞) representations are autoregressive and autoregressive moving average models. In the following sections we prove this.

[1] Note that late on we show that all second order stationary time series $\{X_t\}$ have the representation

$$X_t = \sum_{j=1}^{\infty} \psi_j Z_{t-j}, \tag{4.4}$$

where $\{Z_t = X_t - P_{X_{t-1},X_{t-2},\ldots}(X_t)\}$ (where $P_{X_{t-1},X_{t-2},\ldots}(X_t)$ is the best linear predictor of X_t given the past, $X_{t-1}, X_{t-2}, \ldots$). In this case $\{Z_t\}$ are uncorrelated random variables. It is called Wold's representation theorem (see Section 7.12). The representation in (4.4) has many practical advantages. For example Krampe et al. (2016) recently used it to define the so called "MA bootstrap".

4.3 The AR(p) model

In this section we will examine under what conditions the AR(p) model has a stationary solution.

4.3.1 Difference equations and back-shift operators

The autoregressive model is defined in terms of inhomogenuous difference equations. Difference equations can often be represented in terms of backshift operators, so we start by defining them and see why this representation may be useful (and why it should work).

The time series $\{X_t\}$ is said to be an autoregressive (AR(p)) if it satisfies the equation

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = \varepsilon_t, \quad t \in \mathbb{Z}, \tag{4.5}$$

where $\{\varepsilon_t\}$ are zero mean, finite variance random variables. As we mentioned previously, the autoregressive model is a system of difference equation (which can be treated as a infinite number of simultaneous equations). For this system to make any sense it must have a solution.

Remark 4.3.1 (What is meant by a solution?) *By solution, we mean a sequence of numbers $\{x_t\}_{t=-\infty}^{\infty}$ which satisfy the equations in (7.31). It is tempting to treat (7.31) as a recursion, where we start with an intial value x_I some time far back in the past and use (7.31) to generate $\{x_t\}$ (for a given sequence $\{\varepsilon_t\}_t$). This is true for some equations but not all. To find out which, we need to obtain the solution to (7.31).*

Example Let us suppose the model is

$$X_t = \phi X_{t-1} + \varepsilon_t \textit{ for } t \in \mathbb{Z},$$

where ε_t are iid random variables and ϕ is a known parameter. Let $\varepsilon_2 = 0.5$, $\varepsilon_3 = 3.1$, $\varepsilon_4 = -1.2$ etc. This gives the system of equations

$$x_2 = \phi x_1 + 0.5, \quad x_3 = \phi x_2 + 3.1, \quad \textit{and} \quad x_4 = \phi x_3 - 1.2$$

and so forth. We see this is an equation in terms of unknown $\{x_t\}_t$. Does there exist a $\{x_t\}_t$ which satisfy this system of equations? For linear systems, the answer can easily be found. But more complex systems the answer is not so clear. Our focus in this chapter is on linear systems.

To obtain a solution we write the autoregressive model in terms of backshift operators:

$$X_t - \phi_1 B X_t - \ldots - \phi_p B^p X_t = \varepsilon_t, \quad \Rightarrow \quad \phi(B)X_t = \varepsilon_t$$

where $\phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$, B is the backshift operator and is defined such that $B^k X_t = X_{t-k}$. Simply rearranging $\phi(B)X_t = \varepsilon_t$, gives the 'solution' of the autoregressive difference equation to be $X_t = \phi(B)^{-1}\varepsilon_t$, however this is just an algebraic manipulation, below we investigate whether it really has any meaning.

In the subsections below we will show:

- Let $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j$ be a pth order polynomial in z. Let $z_1, \ldots, z_p$ denote the p roots of $\phi(z)$. A solution for (7.31) will always exist if none of the p roots of $\phi(z)$ lie on the unit circle i.e. $|z_j| \neq 1$ for $1 \leq j \leq p$.

- If all the roots lie outside the unit circle i.e. $|z_j| > 1$ for $1 \leq j \leq p$, then $\{x_t\}$ can be generated by starting with an initial value far in the past x_I and treating (7.31) as a recursion

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \varepsilon_t.$$

  A time series that can be generated using the above recursion is called causal. It will have a very specific solution.

- If all the roots lie inside the unit circle i.e. $|z_j| < 1$ for $1 \leq j \leq p$, then we cannot directly treat (7.31) as a recursion. Instead, we need to rearrange (7.31) such that X_{t-p} is written in terms of $\{X_{t-j}\}_{j=1}^{p}$ and ε_t

$$X_{t-p} = \phi_p^{-1}\left[-\phi_{p-1}X_{t-p+1} - \ldots - \phi_1 X_{t-1} + X_t\right] - \phi_p^{-1}\varepsilon_t. \tag{4.4}$$

  $\{x_t\}$ can be generated by starting with an initial value far in the past x_I and treating (7.31) as a recursion.

- If the roots lie both inside and outside the unit circle. No recursion will generate a solution. But we will show that a solution can be generated by adding recursions together.

To do this, we start with an example.

4.3.2 Solution of two particular AR(1) models

Below we consider two different AR(1) models and obtain their solutions.

(i) Consider the AR(1) process

$$X_t = 0.5X_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z}. \tag{4.5}$$

Notice this is an equation (rather like $3x^2+2x+1 = 0$, or an infinite number of simultaneous equations), which may or may not have a solution. To obtain the solution we note that $X_t = 0.5X_{t-1} + \varepsilon_t$ and $X_{t-1} = 0.5X_{t-2} + \varepsilon_{t-1}$. Using this we get $X_t = \varepsilon_t + 0.5(0.5X_{t-2} + \varepsilon_{t-1}) = \varepsilon_t + 0.5\varepsilon_{t-1} + 0.5^2X_{t-2}$. Continuing this backward iteration we obtain at the kth iteration, $X_t = \sum_{j=0}^{k}(0.5)^j\varepsilon_{t-j} + (0.5)^{k+1}X_{t-k}$. Because $(0.5)^{k+1} \to 0$ as $k \to \infty$ by taking the limit we can show that $X_t = \sum_{j=0}^{\infty}(0.5)^j\varepsilon_{t-j}$ is almost surely finite and a solution of (4.5). Of course like any other equation one may wonder whether it is the unique solution (recalling that $3x^2 + 2x + 1 = 0$ has two solutions). We show in Section 4.3.2 that this is the unique stationary solution of (4.5).

Let us see whether we can obtain a solution using the difference equation representation. We recall, that by crudely taking inverses, the solution is $X_t = (1 - 0.5B)^{-1}\varepsilon_t$. The obvious question is whether this has any meaning. Note that $(1 - 0.5B)^{-1} = \sum_{j=0}^{\infty}(0.5B)^j$, for $|B| \leq 2$, hence substituting this power series expansion into X_t we have

$$X_t = (1 - 0.5B)^{-1}\varepsilon_t = (\sum_{j=0}(0.5B)^j)\varepsilon_t = (\sum_{j=0}(0.5^jB^j))\varepsilon_t = \sum_{j=0}^{\infty}(0.5)^j\varepsilon_{t-j},$$

which corresponds to the solution above. Hence the backshift operator in this example helps us to obtain a solution. Moreover, because the solution can be written in terms of past values of ε_t, it is causal.

(ii) Let us consider the AR model, which we will see has a very different solution:

$$X_t = 2X_{t-1} + \varepsilon_t. \tag{4.6}$$

Doing what we did in (i) we find that after the kth back iteration we have $X_t = \sum_{j=0}^{k} 2^j\varepsilon_{t-j} + 2^{k+1}X_{t-k}$. However, unlike example (i) 2^k does not converge as $k \to \infty$. This suggest that if we continue the iteration $X_t = \sum_{j=0}^{\infty} 2^j \varepsilon_{t-j}$ is not a quantity that is finite (when ε_t are iid). Therefore $X_t = \sum_{j=0}^{\infty} 2^j \varepsilon_{t-j}$ cannot be considered as a solution of (4.6). We need to write (4.6) in a slightly different way in order to obtain a meaningful solution.

Rewriting (4.6) we have $X_{t-1} = 0.5X_t - 0.5\varepsilon_t$. Forward iterating this we get $X_{t-1} = -(0.5)\sum_{j=0}^{k}(0.5)^j \varepsilon_{t+j} - (0.5)^{k+1}X_{t+k}$. Since $(0.5)^{k+1} \to 0$ as $k \to \infty$ we have

$$X_{t-1} = -(0.5)\sum_{j=0}^{\infty}(0.5)^j \varepsilon_{t+j}$$

as a solution of (4.6).

Let us see whether the difference equation can also offer a solution. Since $(1-2B)X_t = \varepsilon_t$, using the crude manipulation we have $X_t = (1-2B)^{-1}\varepsilon_t$. Now we see that

$$(1-2B)^{-1} = \sum_{j=0}^{\infty}(2B)^j \quad \text{for } |B| < 1/2.$$

Using this expansion gives the solution $X_t = \sum_{j=0}^{\infty} 2^j B^j X_t$, but as pointed out above this sum is not well defined. What we find is that $\phi(B)^{-1}\varepsilon_t$ only makes sense (is well defined) if the series expansion of $\phi(B)^{-1}$ converges in a region that includes the unit circle $|B| = 1$.

What we need is another series expansion of $(1-2B)^{-1}$ which converges in a region which includes the unit circle $|B| = 1$ (as an aside, we note that a function does not necessarily have a unique series expansion, it can have difference series expansions which may converge in different regions). We now show that a convergent series expansion needs to be defined in terms of negative powers of B not positive powers. Writing $(1-2B) = -(2B)(1-(2B)^{-1})$, therefore

$$(1-2B)^{-1} = -(2B)^{-1}\sum_{j=0}^{\infty}(2B)^{-j},$$

which converges for $|B| > 1/2$. Using this expansion we have

$$X_t = -\sum_{j=0}^{\infty}(0.5)^{j+1}B^{-j-1}\varepsilon_t = -\sum_{j=0}^{\infty}(0.5)^{j+1}\varepsilon_{t+j+1},$$

which we have shown above is a well defined solution of (4.6).

In summary $(1-2B)^{-1}$ has two series expansions

$$\frac{1}{(1-2B)} = \sum_{j=0}^{\infty}(2B)^{-j}$$

which converges for $|B| < 1/2$ and

$$\frac{1}{(1-2B)} = -(2B)^{-1}\sum_{j=0}^{\infty}(2B)^{-j},$$

which converges for $|B| > 1/2$. The one that is useful for us is the series which converges when $|B| = 1$.

It is clear from the above examples how to obtain the solution of a general AR(1). This solution is unique and we show this below.

Exercise 4.1 *(i) Find the stationary solution of the AR(1) model*

$$X_t = 0.8X_{t-1} + \varepsilon_t$$

where ε_t are iid random variables with mean zero and variance one.

(ii) Find the stationary solution of the AR(1) model

$$X_t = \frac{5}{4}X_{t-1} + \varepsilon_t$$

where ε_t are iid random variables with mean zero and variance one.

(iii) [Optional] Obtain the autocovariance function of the stationary solution for both the models in (i) and (ii).

Uniqueness of the stationary solution the AR(1) model (advanced)

Consider the AR(1) process $X_t = \phi X_{t-1} + \varepsilon_t$, where $|\phi| < 1$. Using the method outlined in (i), it is straightforward to show that $X_t = \sum_{j=0}^{\infty}\phi^j\varepsilon_{t-j}$ is its stationary solution, we now show that this solution is unique. This may seem obvious, but recall that many equations have multiple solutions. The techniques used here generalize to nonlinear models too.

We first show that $X_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$ is well defined (that it is almost surely finite). We note that $|X_t| \leq \sum_{j=0}^{\infty} |\phi^j| \cdot |\varepsilon_{t-j}|$. Thus we will show that $\sum_{j=0}^{\infty} |\phi^j| \cdot |\varepsilon_{t-j}|$ is almost surely finite, which will imply that X_t is almost surely finite. By montone convergence we can exchange sum and expectation and we have $\mathrm{E}(|X_t|) \leq \mathrm{E}(\lim_{n\to\infty} \sum_{j=0}^{n} |\phi^j \varepsilon_{t-j}|) = \lim_{n\to\infty} \sum_{j=0}^{n} |\phi^j| \mathrm{E}|\varepsilon_{t-j}|) = \mathrm{E}(|\varepsilon_0|) \sum_{j=0}^{\infty} |\phi^j| < \infty$. Therefore since $\mathrm{E}|X_t| < \infty$, $\sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$ is a well defined solution of $X_t = \phi X_{t-1} + \varepsilon_t$.

To show that it is the unique, stationary, causal solution, let us suppose there is another (causal) solution, call it Y_t. Clearly, by recursively applying the difference equation to Y_t, for every s we have

$$Y_t = \sum_{j=0}^{s} \phi^j \varepsilon_{t-j} + \phi^s Y_{t-s-1}.$$

Evaluating the difference between the two solutions gives $Y_t - X_t = A_s - B_s$ where $A_s = \phi^s Y_{t-s-1}$ and $B_s = \sum_{j=s+1}^{\infty} \phi^j \varepsilon_{t-j}$ for all s. To show that Y_t and X_t coincide almost surely we will show that for every $\epsilon > 0$, $\sum_{s=1}^{\infty} P(|A_s - B_s| > \varepsilon) < \infty$ (and then apply the Borel-Cantelli lemma). We note if $|A_s - B_s| > \varepsilon)$, then either $|A_s| > \varepsilon/2$ or $|B_s| > \varepsilon/2$. Therefore $P(|A_s - B_s| > \varepsilon) \leq P(|A_s| > \varepsilon/2) + P(|B_s| > \varepsilon/2)$. To bound these two terms we use Markov's inequality. It is straightforward to show that $P(|B_s| > \varepsilon/2) \leq C\phi^s/\varepsilon$. To bound $\mathrm{E}|A_s|$, we note that $|Y_s| \leq |\phi| \cdot |Y_{s-1}| + |\varepsilon_s|$, since $\{Y_t\}$ is a stationary solution then $\mathrm{E}|Y_s|(1 - |\phi|) \leq \mathrm{E}|\varepsilon_s|$, thus $\mathrm{E}|Y_t| \leq \mathrm{E}|\varepsilon_t|/(1 - |\phi|) < \infty$. Altogether this gives $P(|A_s - B_s| > \varepsilon) \leq C\phi^s/\varepsilon$ (for some finite constant C). Hence $\sum_{s=1}^{\infty} P(|A_s - B_s| > \varepsilon) < \sum_{s=1}^{\infty} C\phi^s/\varepsilon < \infty$. Thus by the Borel-Cantelli lemma, this implies that the event $\{|A_s - B_s| > \varepsilon\}$ happens only finitely often (almost surely). Since for every ε, $\{|A_s - B_s| > \varepsilon\}$ occurs (almost surely) only finitely often for all ε, then $Y_t = X_t$ almost surely. Hence $X_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$ is (almost surely) the unique causal solution.

4.3.3 The solution of a general AR(p)

Let us now summarise our observation for the general AR(1) process $X_t = \phi X_{t-1} + \varepsilon_t$. If $|\phi| < 1$, then the solution is in terms of past values of $\{\varepsilon_t\}$, if on the other hand $|\phi| > 1$ the solution is in terms of future values of $\{\varepsilon_t\}$.

In this section we focus on general AR(p) model

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = \varepsilon_t, \quad t \in \mathbb{Z}, \tag{4.7}$$

Generalising this argument to a general polynomial, if the roots of $\phi(B)$ are greater than one, then the power series of $\phi(B)^{-1}$ (which converges for $|B| = 1$) is in terms of positive powers (hence the solution $\phi(B)^{-1}\varepsilon_t$ will be in past terms of $\{\varepsilon_t\}$). On the other hand, if the roots are both less than and greater than one (but do not lie on the unit circle), then the power series of $\phi(B)^{-1}$ will be in both negative and positive powers. Thus the solution $X_t = \phi(B)^{-1}\varepsilon_t$ will be in terms of both past and future values of $\{\varepsilon_t\}$. We summarize this result in a lemma below.

Lemma 4.3.1 *Suppose that the AR(p) process satisfies the representation $\phi(B)X_t = \varepsilon_t$, where none of the roots of the characteristic polynomial lie on the unit circle and $\mathrm{E}|\varepsilon_t| < \infty$. Then $\{X_t\}$ has a stationary, almost surely unique, solution*

$$X_t = \sum_{j\in\mathbb{Z}} \psi_j \varepsilon_{t-j}$$

where $\psi(z) = \sum_{j\in\mathbb{Z}} \psi_j z^j = \phi(z)^{-1}$ (the Laurent series of $\phi(z)^{-1}$ which converges when $|z| = 1$).

We see that where the roots of the characteristic polynomial $\phi(B)$ lie defines the solution of the AR process. We will show in Sections **??** and 6.1.2 that it not only defines the solution but also determines some of the characteristics of the time series.

Exercise 4.2 *Suppose $\{X_t\}$ satisfies the AR(p) representation*

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t,$$

where $\sum_{j=1}^{p} |\phi_j| < 1$ and $\mathrm{E}|\varepsilon_t| < \infty$. Show that $\{X_t\}$ will always have a causal stationary solution (i.e. the roots of the characteristic polynomial are outside the unit circle).

4.3.4 Obtaining an explicit solution of an AR(2) model

A worked out example

Suppose $\{X_t\}$ satisfies

$$X_t = 0.75X_{t-1} - 0.125X_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables. We want to obtain a solution for the above equations.

It is not easy to use the backward (or forward) iterating techique for AR processes beyond order one. This is where using the backshift operator becomes useful. We start by writing $X_t = 0.75X_{t-1} - 0.125X_{t-2} + \varepsilon_t$ as $\phi(B)X_t = \varepsilon$, where $\phi(B) = 1 - 0.75B + 0.125B^2$, which leads to what is commonly known as the characteristic polynomial $\phi(z) = 1 - 0.75z + 0.125z^2$. If we can find a power series expansion of $\phi(B)^{-1}$, which is valid for $|B| = 1$, then the solution is $X_t = \phi(B)^{-1}\varepsilon_t$.

We first observe that $\phi(z) = 1 - 0.75z + 0.125z^2 = (1 - 0.5z)(1 - 0.25z)$. Therefore by using partial fractions we have

$$\frac{1}{\phi(z)} = \frac{1}{(1-0.5z)(1-0.25z)} = \frac{-1}{(1-0.5z)} + \frac{2}{(1-0.25z)}.$$

We recall from geometric expansions that

$$\frac{-1}{(1-0.5z)} = -\sum_{j=0}^{\infty}(0.5)^j z^j \quad |z| \leq 2, \qquad \frac{2}{(1-0.25z)} = 2\sum_{j=0}^{\infty}(0.25)^j z^j \quad |z| \leq 4.$$

Putting the above together gives

$$\frac{1}{(1-0.5z)(1-0.25z)} = \sum_{j=0}^{\infty}\{-(0.5)^j + 2(0.25)^j\}z^j \quad |z| < 2.$$

The above expansion is valid for $|z| = 1$, because $\sum_{j=0}^{\infty}| - (0.5)^j + 2(0.25)^j| < \infty$ (see Lemma 4.3.2). Hence

$$X_t = \{(1-0.5B)(1-0.25B)\}^{-1}\varepsilon_t = \big(\sum_{j=0}^{\infty}\{-(0.5)^j + 2(0.25)^j\}B^j\big)\varepsilon_t = \sum_{j=0}^{\infty}\{-(0.5)^j + 2(0.25)^j\}\varepsilon_{t-j},$$

which gives a stationary solution to the AR(2) process (see Lemma 4.2.1). Moreover since the roots lie outside the unit circle the solution is *causal*.

The discussion above shows how the backshift operator can be applied and how it can be used to obtain solutions to AR(p) processes.

The solution of a general AR(2) model

We now generalise the above to general AR(2) models

$$X_t = (a+b)X_{t-1} - abX_{t-2} + \varepsilon_t,$$

the characteristic polynomial of the above is $1-(a+b)z+abz^2 = (1-az)(1-bz)$. This means the solution of X_t is

$$X_t = (1-Ba)^{-1}(1-Bb)^{-1}\varepsilon_t,$$

thus we need an expansion of $(1-Ba)^{-1}(1-Bb)^{-1}$. Assuming that $a \neq b$, and using partial fractions we have

$$\frac{1}{(1-za)(1-zb)} = \frac{1}{b-a}\left(\frac{b}{1-bz} - \frac{a}{1-az}\right)$$

Cases:

(1) $|a| < 1$ and $|b| < 1$, this means the roots lie outside the unit circle. Thus the expansion is

$$\frac{1}{(1-za)(1-zb)} = \frac{1}{(b-a)}\big(b\sum_{j=0}^{\infty} b^j z^j - a\sum_{j=0}^{\infty} a^j z^j\big),$$

which leads to the causal solution

$$X_t = \frac{1}{b-a}\Big(\sum_{j=0}^{\infty}\big(b^{j+1} - a^{j+1}\big)\varepsilon_{t-j}\Big). \tag{4.8}$$

(2) Case that $|a| > 1$ and $|b| < 1$, this means the roots lie inside and outside the unit circle and we have the expansion

$$\begin{aligned}\frac{1}{(1-za)(1-zb)} &= \frac{1}{b-a}\left(\frac{b}{1-bz} - \frac{a}{(az)((az)^{-1}-1)}\right)\\ &= \frac{1}{(b-a)}\left(b\sum_{j=0}^{\infty} b^j z^j + z^{-1}\sum_{j=0}^{\infty} a^{-j}z^{-j}\right),\end{aligned} \tag{4.9}$$

which leads to the non-causal solution

$$X_t = \frac{1}{b-a}\big(\sum_{j=0}^{\infty} b^{j+1}\varepsilon_{t-j} + \sum_{j=0}^{\infty} a^{-j}\varepsilon_{t+1+j}\big). \tag{4.10}$$

[2]

[2] Later we show that the non-causal X_t, has the same correlation as an AR(2) model whose characteristic polynomial has the roots a^{-1} and b, since both these roots lie out side the unit this model has a causal solution. Moreover, it is possible to rewrite this non-causal AR(2) as an MA infinite type process but where

Returning to (4.10), we see that this solution throws up additional interesting results. Let us return to the expansion in (4.9) and apply it to X_t

$$\begin{aligned} X_t &= \frac{1}{(1-Ba)(1-Bb)}\varepsilon_t = \frac{1}{b-a}\left(\underbrace{\frac{b}{1-bB}\varepsilon_t}_{\text{causal AR(1)}} + \underbrace{\frac{1}{B(1-a^{-1}B^{-1})}\varepsilon_t}_{\text{noncausal AR(1)}}\right) \\ &= \frac{1}{b-a}\left(Y_t + Z_{t+1}\right) \end{aligned}$$

where $Y_t = bY_{t-1} + \varepsilon_t$ and $Z_{t+1} = a^{-1}Z_{t+2} + \varepsilon_{t+1}$. In other words, the noncausal AR(2) process is the sum of a causal and a'future' AR(1) process. This is true for all noncausal time series (except when there is multiplicity in the roots) and is discussed further in Section **??**.

We mention that several authors argue that noncausal time series can model features in data which causal time series cannot.

(iii) $a = b < 1$ (both roots are the same and lie outside the unit circle). The characteristic polynomial is $(1-az)^2$. To obtain the convergent expansion when $|z| = 1$ we note that $(1-az)^{-2} = (-1)\frac{d(1-az)^{-1}}{d(az)}$. Thus

$$\frac{(-1)}{(1-az)^2} = (-1)\sum_{j=0}^{\infty} j(az)^{j-1}.$$

This leads to the causal solution

$$X_t = (-1)\sum_{j=1}^{\infty} ja^{j-1}\varepsilon_{t-j}.$$

In many respects this is analogous to Matern covariance defined over $\mathbb{R}^d$ (and used in spatial statistics). However, unlike autocovarianced defined over $\mathbb{R}^d$ the behaviour of the autocovari-

the innovations are no independent but uncorrelated instead. I.e. we can write X_t as

$$(1-a^{-1}B)(1-bB)X_t = \widetilde{\varepsilon}_t,$$

where $\widehat{\varepsilon}_t$ are uncorrelated (and are a linear sum of the iid $varepsilon_t$), which as the solution

$$X_t = \frac{1}{b-a}\left(\sum_{j=0}^{\infty}\left(b^{j+1} - a^{j+1}\right)\widetilde{\varepsilon}_{t-j}\right). \tag{4.11}$$

ance at zero is not an issue.

Exercise 4.3 *Show for the $AR(2)$ model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t$ to have a causal stationary solution the parameters ϕ_1, ϕ_2 must lie in the region defined by the three conditions*

$$\phi_2 + \phi_1 < 1, \quad \phi_2 - \phi_1 < 1 \quad |\phi_2| < 1.$$

Exercise 4.4 *(a) Consider the $AR(2)$ process*

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance one. Suppose the absolute of the roots of the characteristic polynomial $1 - \phi_1 z - \phi_2 z^2$ are greater than one. Show that $|\phi_1| + |\phi_2| < 4$.

(b) Now consider a generalisation of this result. Consider the $AR(p)$ process

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots \phi_p X_{t-p} + \varepsilon_t.$$

Suppose the absolute of the roots of the characteristic polynomial $1 - \phi_1 z - \ldots - \phi_p z^p$ are greater than one. Show that $|\phi_1| + \ldots + |\phi_p| \leq 2^p$.

4.3.5 History of the periodogram (Part II)

We now return to the development of the periodogram and the role that the AR model played in understanding its behaviour.

The general view until the 1920s is that most time series were a mix of periodic function with additive noise (where we treat Y_t as the yearly sunspot data)

$$Y_t = \sum_{j=1}^{P} [A_j \cos(t\Omega_j) + B_j \sin(t\Omega_j)] + \varepsilon_t.$$

In the 1920's, Udny Yule, a statistician, and Gilbert Walker, a Meterologist (working in Pune, India) believed an alternative model could be used to explain the features seen in the periodogram. Yule fitted an Autoregressive model of order two to the Sunspot data and obtained the AR(2) model

$$X_t = 1.381X_{t-1} - 0.6807X_{t-2} + \varepsilon_t.$$

We simulate a Gaussian model with exactly this AR(2) structure. In Figure 4.2 plot of the sunspot data together realisation of the AR(2) process. In Figure 4.1 we plot the periodogram of the sunspot data and a realisation from the fitted AR(2) process. One can fit a model to any data set. What makes this model so interesting, is that the simple AR(2) models, model suprisingly well many of the prominent features seen in the sunspot data. From Figures 4.1 and 4.2 we see how well the AR(2) which is full stochastic can model a periodicities.

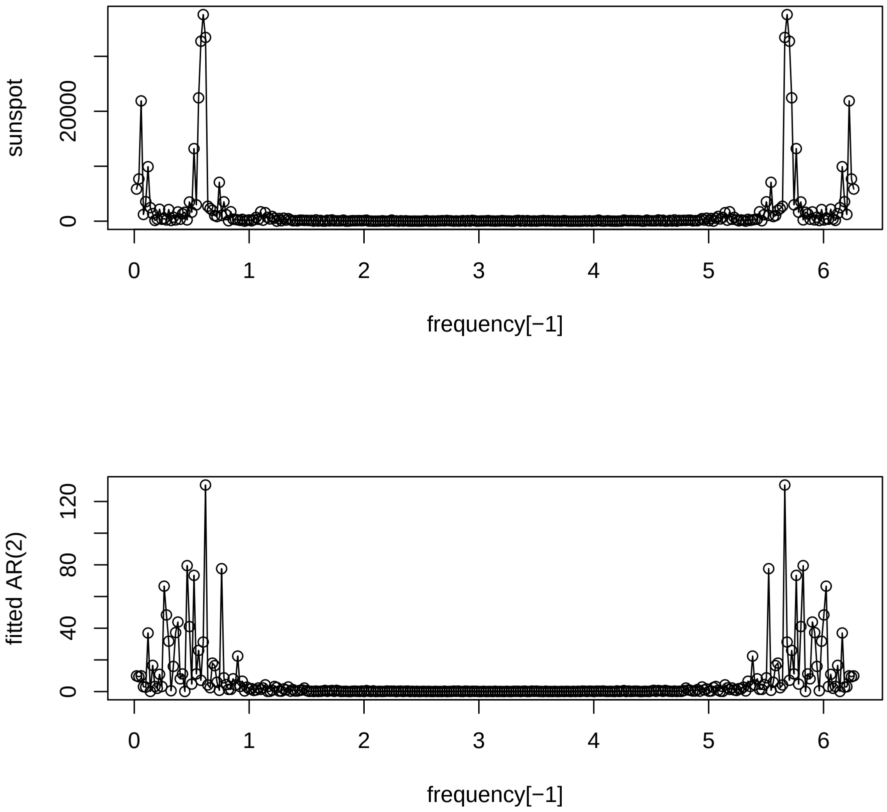


Figure 4.1: The periodogram of the Sunspot data is the top plot and the periodogram of the fitted AR(2) model is the lower plot. They do not look exactly the same, but the AR(2) model is able to capture some of the periodicities.

To summarize, Schuster, and Yule and Walker fit two completely different models to the same

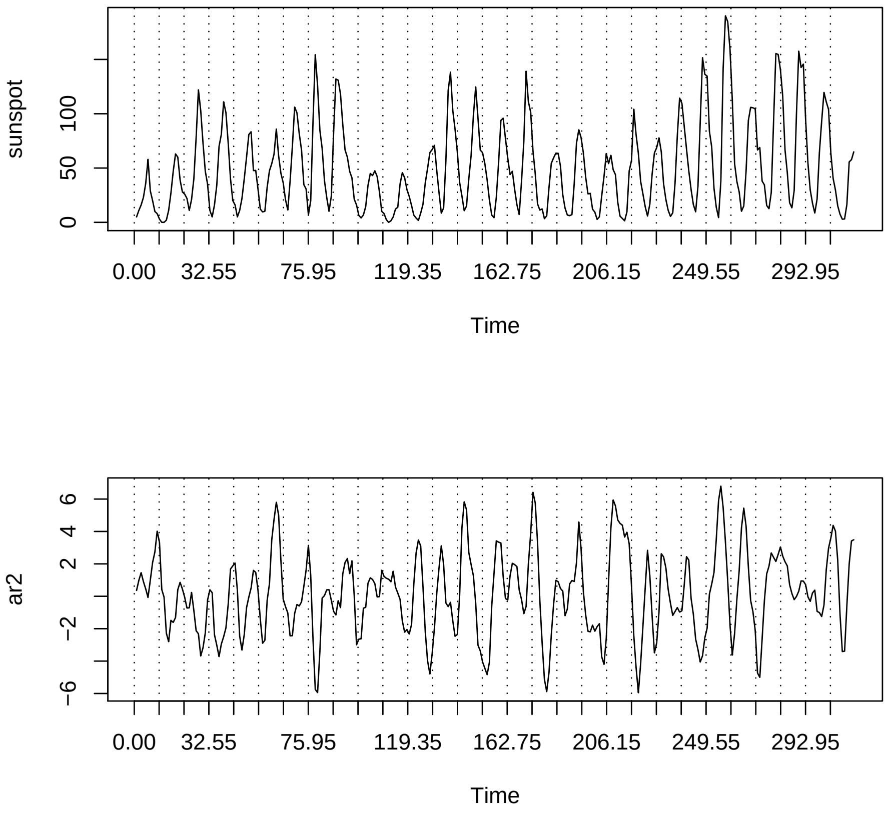


Figure 4.2: Top: Sunspot, Lower: a realisation from the AR(2) process. Lines correspond to period of $P = 2\pi/0.57 = 10.85$ years.

data set and both models are able to mimic the periodocities observed in the sunspot data. While it is obvious how a superimposition of sines and cosines can model periodicities it is not so clear how the AR(2) can achieve a similar effect.

In the following section we study the coefficient of the AR(2) model and how it can mimic the periodicities seen in the data.

4.3.6 Examples of "Pseudo" periodic AR(2) models

We start by studying the AR(2) model that Yule and Walker fitted to the data. We recall that the fitted coefficients were

$$X_t = 1.381X_{t-1} - 0.6807X_{t-2} + \varepsilon_t.$$

This corresponds to the characteristic function $\phi(z) = 1 - 1.381z + 0.68z^2$. The roots of this polynomial are $\lambda_1 = 0.77^{-1}\exp(i0.57)$ and $\lambda_2 = 077^{-1}\exp(-i0.57)$. Cross referencing with the periodogram in Figure 4.1, we observe that the peak in the periodogram is at around 0.57 also. This suggests that the phase of the solution (in polar form) determines the periodicities. If the solution is real then the phase is either 0 or π and X_t has no (pseudo) periodicities or alternates between signs.

Observe that complex solutions of $\phi(z)$ must have conjugations in order to ensure $\phi(z)$ is real. Thus if a solution of the characteristic function corresponding to an AR(2) is $\lambda_1 = r\exp(i\theta)$, then $\lambda_2 = r\exp(-i\theta)$. Based on this $\phi(z)$ can be written as

$$\phi(z) = (1 - r\exp(i\theta)z)(1 - r\exp(-i\theta)) = 1 - 2r\cos(\theta)z + r^2z^2,$$

this leads to the AR(2) model

$$X_t = 2r\cos(\theta)X_{t-1} - r^2X_{t-2} + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid random variables. To ensure it is causal we set $|r| < 1$. In the simulations below we consider the models

$$X_t = 2r\cos(\pi/3)X_{t-1} - r^2X_{t-2} + \varepsilon_t$$

and

$$X_t = 2r\cos(0)X_{t-1} - r^2X_{t-2} + \varepsilon_t$$

for $r = 0.5$ and $r = 0.9$.The latter model has completely real coefficients and its characteristic function is $\phi(z) = (1 - rz)^2$.

In Figures 4.3 and 4.4 we plot a typical realisation from these models with $n = 200$ and corresponding periodogram for the case $\theta = \pi/3$. In Figures 4.5 and 4.6 we plot the a typical realisation and corresponding periodogram for the case $\theta = 0$

From the realisations and the periodogram we observe a periodicity centered about frequency $\pi/3$ or 0 (depending on the model). We also observe that the larger r is the more pronounced the period. For frequency 0, there is no period it is simply what looks like trend (very low frequency

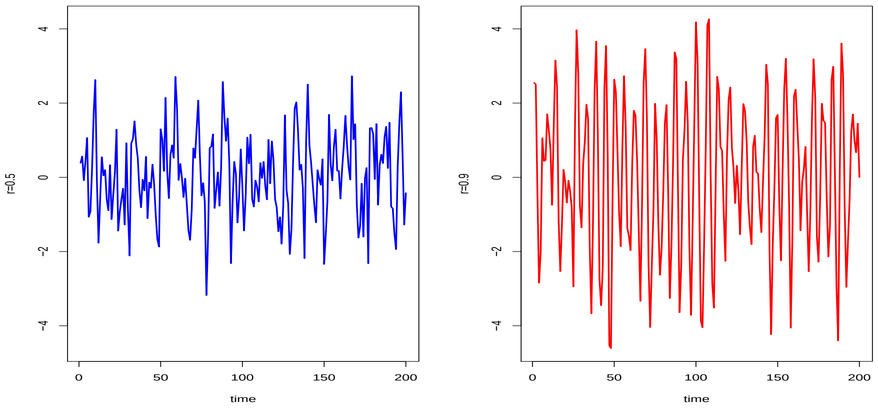


Figure 4.3: Realisation for $X_t = 2r\cos(\pi/3)X_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

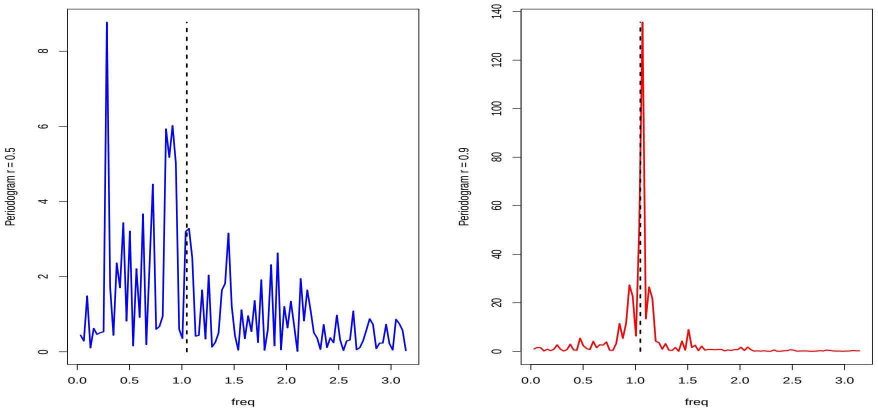


Figure 4.4: Periodogram for realisation from $X_t = 2r\cos(\pi/3)X_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

behaviour). But the AR(2) is a completely stochastic system (random), it is strange that exhibits behaviour close to period. We explain why in the following section.

We conclude this section by showing what shape the periodogram is trying to mimic (but not so well!). In will be shown later on that the expectation of the peridodogram is roughly equal to the spectral density function of the AR(2) process which is

$$f(\omega) = \frac{1}{|1 - \phi_1 e^{i\omega} - \phi_2 e^{i2\omega}|^2} = \frac{1}{|1 - 2r\cos\theta e^{i\omega} + r^2 e^{2i\omega}|^2}.$$

Plots of the spectral density for $\theta = \pi/3$, $\theta = 0$ and $r = 0.5$ and 0.9 are given in Figures 4.7 and 4.8. Observe that the shapes in Figures 4.4 and 4.6 match those in Figures 4.7 and 4.8. But the

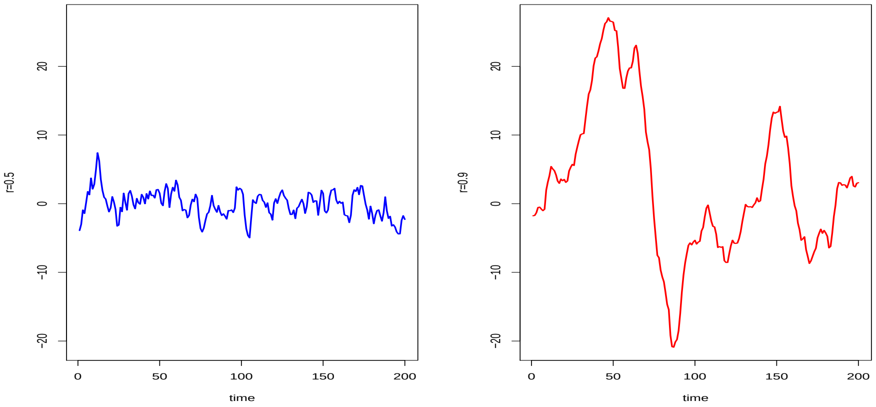


Figure 4.5: Realisation for $X_t = 2rX_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

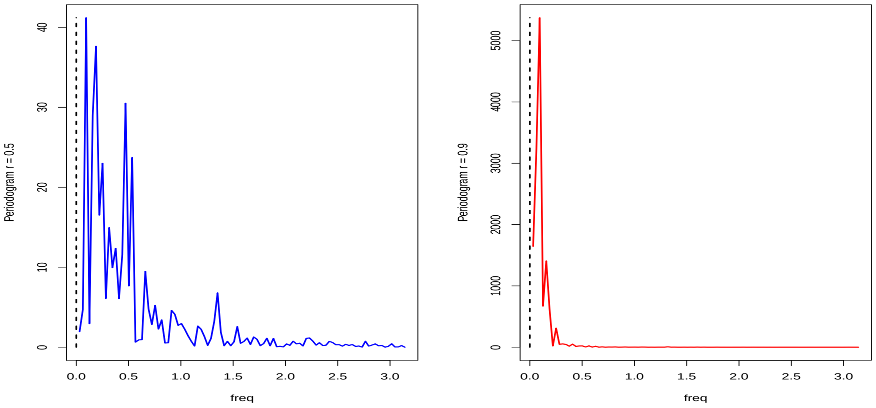


Figure 4.6: Periodogram for realisation from $X_t = 2rX_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

periodogram is very rough whereas the spectral density is smooth. This is because the periodogram is simply a mirror of all the frequencies in the observed time series, and the actual time series do not contain any pure frequencies. It is a mismatch of cosines and sines, thus the messiness of the periodogram.

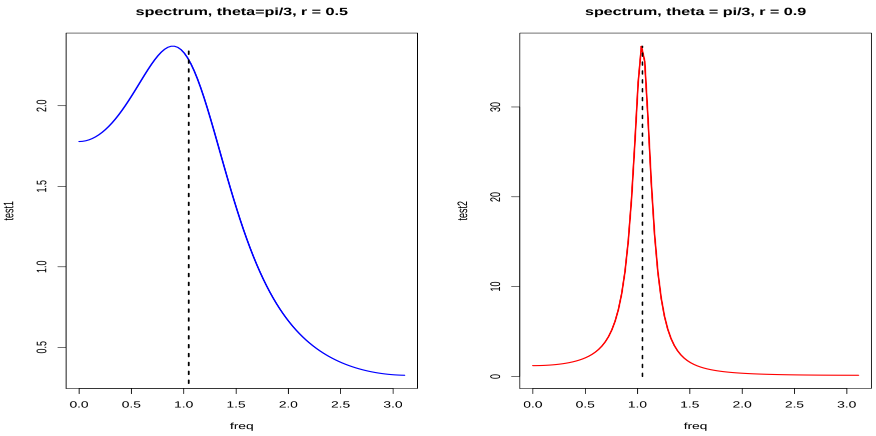


Figure 4.7: Spectral density for $X_t = 2rcos(\pi/3)X_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

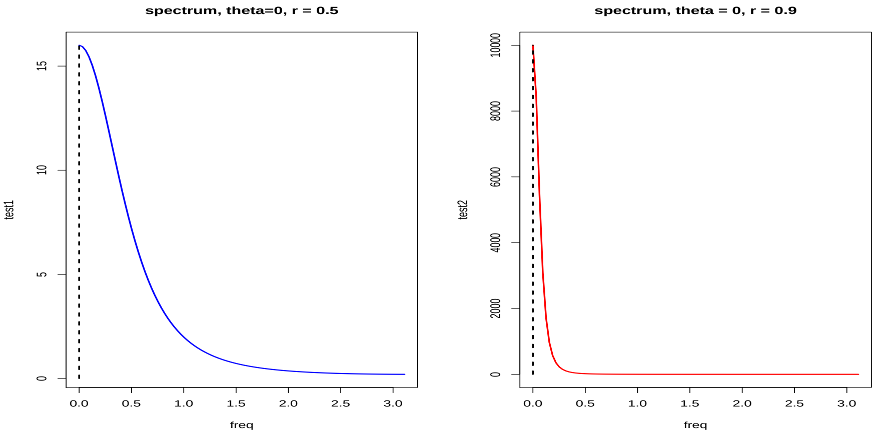


Figure 4.8: Spectral density for $X_t = 2rX_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

4.3.7 Derivation of "Pseudo" periodicity functions in an AR(2)

We now explain why the AR(2) (and higher orders) can characterise some very interesting behaviour (over the rather dull AR(1)). For now we assume that X_t is a causal time series which satisfies the AR(2) representation

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid with mean zero and finite variance. We focus on the case that the characteristic polynomial is complex with roots $\lambda_1 = r\exp(i\theta)$ and $\lambda_2 = r\exp(-i\theta)$. Thus our focus is on the

AR(2) model

$$X_t = 2r\cos(\theta)X_{t-1} - r^2 X_{t-2} + \varepsilon_t \qquad |r| < 1.$$

By using equation (4.8) with $a = \lambda$ and $b = \overline{\lambda}$

$$X_t = \frac{1}{\lambda - \overline{\lambda}} \sum_{j=0}^{\infty} \left(\lambda^{j+1} - \overline{\lambda^{j+1}} \right) \varepsilon_{t-j}.$$

We reparameterize $\lambda = re^{i\theta}$ (noting that $|r| < 1$). Then

$$X_t = \frac{1}{2r\sin\theta} \sum_{j=0}^{\infty} 2r^{j+1} \sin\left((j+1)\theta\right) \varepsilon_{t-j}. \tag{4.12}$$

We can see that X_t is effectively the sum of cosines/sines with frequency θ that have been modulated by the iid errors and exponentially damped. This is why for realisations of autoregressive processes you will often see periodicities (depending on the roots of the characteristic). Thus to include periodicities in a time series in an These arguments can be generalised to higher order autoregressive models.

Exercise 4.5 *(a) Obtain the stationary solution of the AR(2) process*

$$X_t = \frac{7}{3}X_{t-1} - \frac{2}{3}X_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2.

Does the solution have an MA(∞) representation?

(b) Obtain the stationary solution of the AR(2) process

$$X_t = \frac{4 \times \sqrt{3}}{5} X_{t-1} - \frac{4^2}{5^2} X_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2.

Does the solution have an MA(∞) representation?

(c) Obtain the stationary solution of the AR(2) process

$$X_t = X_{t-1} - 4X_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2.

Does the solution have an $MA(\infty)$ representation?

Exercise 4.6 *Construct a causal stationary $AR(2)$ process with pseudo-period 17. Using the* `R` *function* `arima.sim` *simulate a realisation from this process (of length 200) and make a plot of the periodogram. What do you observe about the peak in this plot?*

4.3.8 Seasonal Autoregressive models

A popular autoregessive model that is often used for modelling seasonality, is the seasonal autoregressive model (SAR). To motivate the model consider the monthly average temperatures in College Station. Let $\{X_t\}$ denote the monthly temperatures. Now if you have had any experience with temperatures in College Station using the average temperature in October (still hot) to predict the average temperature in November (starts to cool) may not seem reasonable. It may seem more reasonable to use the temperature last November. We can do this using the following model

$$X_t = \phi X_{t-12} + \varepsilon_t,$$

where $|\phi| < 1$. This is an AR(12) model in disguise, The characteristic function $\phi(z) = 1 - \phi z^{12}$ has roots $\lambda_j = \phi^{-1/12}\exp(i2\pi j/12)$ for $j = 0, 1, \ldots, 11$. As there are 5 complex pairs and two real terms. We would expect to see 7 peaks in the periodogram and spectral density. The spectral density is

$$f(\omega) = \frac{1}{|1 - \phi e^{i12\omega}|^2}.$$

A realisation from the above model with $\phi = 0.8$ and $n = 200$ is given in Figure 4.9. The corresponding periodogram and spectral density is given in Figure 4.10. We observe that the periodogram captures the general peaks in the spectral density, but is a lot messier.

4.3.9 Solution of the general AR(∞) model (advanced)

The AR(∞) model generalizes the AR(p)

$$X_t = \sum_{j=1}^{\infty} \phi_j X_{t-j} + \varepsilon_t$$

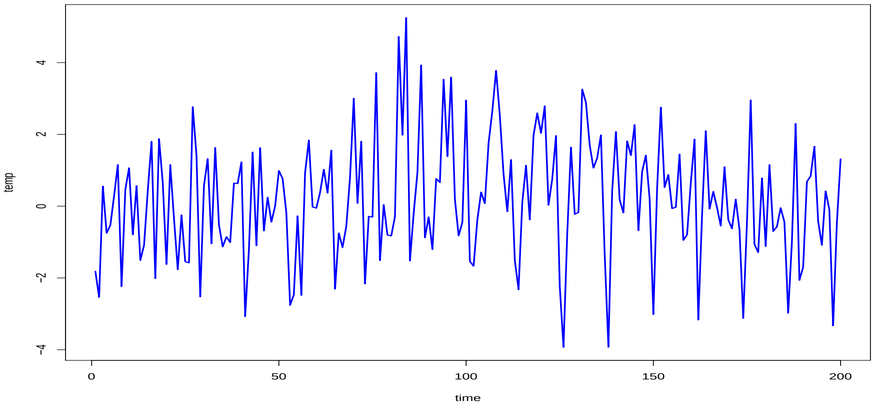


Figure 4.9: Realisation from SAR(12) with $\phi = 0.8$.

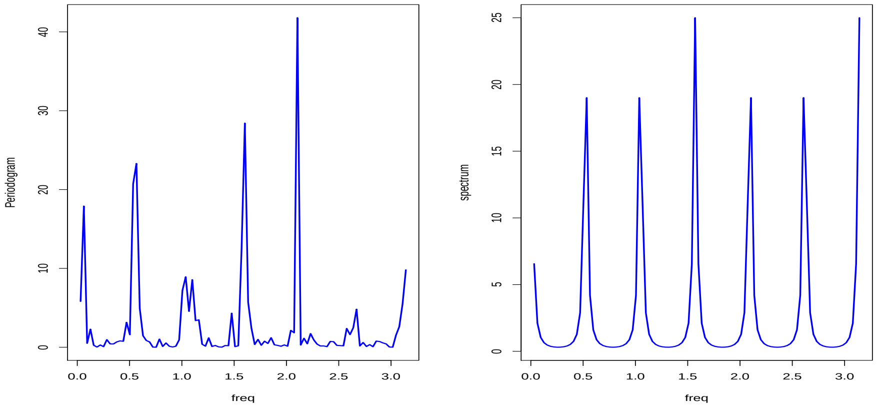


Figure 4.10: Left: Periodogram of realisation. Right Spectral density of model.

where $\{\varepsilon_t\}$ are iid random variables. AR(∞) models are more general than the AR(p) model and are able to model more complex behaviour, such as slower decay of the covariance structure.

In order to obtain the stationary solution of an AR(∞), we need to define an analytic function and its inverse.

Definition 4.3.1 (Analytic functions in the region Ω) *Suppose that $z \in \mathbb{C}$. $\phi(z)$ is an analytic complex function in the region Ω, if it has a power series expansion which converges in Ω, that is $\phi(z) = \sum_{j=-\infty}^{\infty} \phi_j z^j$.*

If there exists a function $\tilde{\phi}(z) = \sum_{j=-\infty}^{\infty} \tilde{\phi}_j z^j$ such that $\tilde{\phi}(z)\phi(z) = 1$ for all $z \in \Omega$, then $\tilde{\phi}(z)$ is the inverse of $\phi(z)$ in the region Ω.

Example 4.3.1 (Analytic functions) *(i) Clearly $a(z) = 1 - 0.5z$ is analytic for all $z \in \mathbb{C}$,* *and has no zeros for $|z| < 2$. The inverse is $\frac{1}{a(z)} = \sum_{j=0}^{\infty}(0.5z)^j$ is well defined in the region $|z| < 2$.*

(ii) Clearly $a(z) = 1 - 2z$ is analytic for all $z \in \mathbb{C}$, and has no zeros for $|z| > 1/2$. The inverse is $\frac{1}{a(z)} = (-2z)^{-1}(1 - (1/2z)) = (-2z)^{-1}(\sum_{j=0}^{\infty}(1/(2z))^j)$ well defined in the region $|z| > 1/2$.

(iii) The function $a(z) = \frac{1}{(1-0.5z)(1-2z)}$ is analytic in the region $0.5 < z < 2$.

(iv) $a(z) = 1 - z$, is analytic for all $z \in \mathbb{C}$, but is zero for $z = 1$. Hence its inverse is not well defined for regions which involve $|z| = 1$ (see Example 4.7).

(v) Finite order polynomials such as $\phi(z) = \sum_{j=0}^{p} \phi_j z^j$ for $\Omega = \mathbb{C}$.

(vi) The expansion $(1 - 0.5z)^{-1} = \sum_{j=0}^{\infty}(0.5z)^j$ for $\Omega = \{z; |z| \leq 2\}$.

We observe that for AR processes we can represent the equation as $\phi(B)X_t = \varepsilon_t$, which formally gives the solution $X_t = \phi(B)^{-1}\varepsilon_t$. This raises the question, under what conditions on $\phi(B)^{-1}$ is $\phi(B)^{-1}\varepsilon_t$ a valid solution. For $\phi(B)^{-1}\varepsilon_t$ to make sense $\phi(B)^{-1}$ should be represented as a power series expansion. Below, we state a technical lemma on $\phi(z)$ which we use to obtain a stationary solution.

Lemma 4.3.2 (Technical lemma) *Suppose that $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$ is finite on a region that includes $|z| = 1$ (we say it is analytic in the region $|z| = 1$). Then $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.*

An immediate consequence of the lemma above is that if $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$ is analytic in the region and $\{X_t\}$ is a strictly stationary time series, where $\mathrm{E}|X_t|$ we define the time series $Y_t = \psi(B)X_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$. Then by the lemma above and Lemma 4.2.1, $\{Y_t\}$ is almost surely finite and strictly stationary time series. We use this result to obtain a solution of an AR(∞) (which includes an AR(p) as a special case).

Lemma 4.3.3 *Suppose $\phi(z) = 1 + \sum_{j=1}^{\infty} \phi_j$ and $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$ are analytic functions in a region which contains $|z| = 1$ and $\phi(z)\psi(z)^{-1} = 1$ for all $|z| = 1$. Then the AR(∞) process*

$$X_t = \sum_{j=1}^{\infty} \phi_j X_{t-j} + \varepsilon_t.$$

has the unique solution

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

We can immediately apply the lemma to find conditions under which the AR(p) process will admit a stationary solution. Note that this is generalisation of Lemma 4.3.1.

Rules of the back shift operator:

(i) If $a(z)$ is analytic in a region Ω which includes the unit circle $|z| = 1$ in its interior and $\{Y_t\}$ is a well defined time series, then X_t defined by $Y_t = a(B)X_t$ is a well defined random variable.

(ii) The operator is commutative and associative, that is $[a(B)b(B)]X_t = a(B)[b(B)X_t] = [b(B)a(B)]X_t$ (the square brackets are used to indicate which parts to multiply first). This may seem obvious, but remember matrices are not commutative!

(iii) Suppose that $a(z)$ and its inverse $\frac{1}{a(z)}$ are both have solutions in the region Ω which includes the unit circle $|z| = 1$ in its interior. If $a(B)X_t = Z_t$, then $X_t = \frac{1}{a(B)}Z_t$.

The magic backshift operator

A precise proof of Lemma 4.3.3 and the rules of the back shift operator described above is beyond these notes. But we briefly describe the idea, so the backshift operator feels less like a magic trick.

Equation (4.7) is an infinite dimension matrix operation that maps (ℓ_2-sequences to ℓ_2-sequences) where $\Gamma : \ell_2 \to \ell_2$ and $\Gamma(x) = \varepsilon$ with $x = (\ldots, x_{-1}, x_0, x_1, \ldots)$. Thus $x = \Gamma^{-1}\varepsilon$. The objectives is to find the coefficients in the operator Γ^{-1}. It is easier to do this by transforming the operator to the Fourier domain with the Fourier operator $F : \ell_2 \to L_2[0, 2\pi]$ and $F^* : L_2[0, 2\pi] \to \ell_2$. Thus $F\Gamma F^*$ is an integral operator with kernel $K(\lambda, \omega) = \phi(e^{i\omega})\delta_{\omega=\lambda}$. It can be shown that the inverse operator $(F\Gamma^{-1}F^*)$ has kernel $K^{-1}(\lambda, \omega) = \phi(e^{i\omega})^{-1}\delta_{\omega=\lambda}$. One can then deduce that the coefficients of Γ^{-1} are the Fourier coefficients $\int_0^{2\pi} \phi(e^{i\omega})^{-1} e^{-ij\omega} d\omega$, which correspond to the expansion of $\phi(z)^{-1}$ that converges in the region that include $|z| = 1$ (the Laurent series in this region).

AR(∞) representation of stationary time series (Advanced)

If a time series is second order stationary and its spectral density function $f(\omega) = (2\pi)^{-1} \sum_{r\in\mathbb{Z}} c(r)e^{ir\omega}$ is bounded away from zero (is not zero) and is finite on $[0, \pi]$. Then it will have form of AR(∞)

representation

$$X_t = \sum_{j=1}^{\infty} a_j X_{t-j} + \varepsilon_t,$$

the difference is that $\{\varepsilon_t\}$ are **uncorrelated random variables** and **may not be** iid random variables. This result is useful when finding the best linear predictors of X_t given the past.

4.4 Simulating from an Autoregressive process

Simulating from a Gaussian AR process

We start with the case that the innovations, $\{\varepsilon_t\}$, are Gaussian. In this case, by using Lemma 4.5.1(ii) we observe that all AR processes can be written as the infinite sum of the innovations. As sums of iid Gaussian random variables are Gaussian, then the resulting time series is also Gaussian. We show in Chapter 6 that given any causal AR equation, the covariance structure of the time series can be deduced. Since normal random variables are fully determined by their mean and variance matrix, using the function `mvnorm` and $\text{var}[\underline{X}_p] = \Sigma_p$, we can simulate the first p elements in the time series $\underline{X}_p = (X_1, \ldots, X_p)$. Then by simulating $(n-p)$ iid random variables we can generate X_t using the causal recursion

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t.$$

Remark 4.4.1 *Any non-causal system of difference equations with Gaussian innovations can always be rewritten as a causal system. This property is unique for Gaussian processes.*

A worked example

We illustrate the details with with an AR(1) process. Suppose $X_t = \phi_1 X_{t-1} + \varepsilon_t$ where $\{\varepsilon_t\}$ are iid standard normal random variables (note that for Gaussian processes it is impossible to discriminate between causal and non-causal processes - see Section 6.4, therefore we will assume $|\phi_1| < 1$). We will show in Section 6.1, equation (6.1) that the autocovariance of an AR(1) is

$$c(r) \;\; = \;\; \phi_1^r \sum_{j=0}^{\infty} \phi_1^{2j} = \frac{\phi_1^r}{1-\phi_1^2}.$$

Therefore, the marginal distribution of X_t is Gaussian with variance $(1-\phi_1^2)^{-1}$. Therefore, to simulate an AR(1) Gaussian time series, we draw from a Gaussian time series with mean zero and variance $(1-\phi_1^2)^{-1}$, calling this X_1. We then iterate for $2 \leq t$, $X_t = \phi_1 X_{t-1} + \varepsilon_t$. This will give us a stationary realization from an AR(1) Gaussian time series.

Note the function `arima.sim` is a routine in `R` which does the above. See below for details.

Simulating from a non-Gaussian causal AR model

Unlike the Gaussian AR process it is difficult to simulate an exact non-Gaussian model, but we can obtain a very close approximation. This is because if the innovations are non-Gaussian the distribution of X_t is not simple. Here we describe how to obtain a close approximation in the case that the AR process is causal.

A worked example We describe a method for simulating an AR(1). Let $\{X_t\}$ be an AR(1) process, $X_t = \phi_1 X_{t-1} + \varepsilon_t$, which has stationary, causal solution

$$X_t = \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}.$$

To simulate from the above model, we set $\tilde{X}_1 = 0$. Then obtain the iteration $\tilde{X}_t = \phi_1 \tilde{X}_{t-1} + \varepsilon_t$ for $t \geq 2$. We note that the solution of this equation is

$$\tilde{X}_t = \sum_{j=0}^{t} \phi_1^j \varepsilon_{t-j}.$$

We recall from Lemma 4.5.1 that $|X_t - \tilde{X}_t| \leq |\phi_1|^t \sum_{j=0}^{\infty} |\phi_1^j \varepsilon_{-j}|$, which converges geometrically fast to zero. Thus if we choose a large n to allow 'burn in' and use $\{\tilde{X}_t; t \geq n\}$ in the simulations we have a simulation which is close to a stationary solution from an AR(1) process. Using the same method one can simulate causal AR(p) models too.

Building AR(p) models One problem with the above approach is the AR(p) coefficients $\{\phi_j\}$ should be chosen such that it corresponds to a causal solution. This is not so simple. It is easier to build a causal AR(p) model from its factorisation:

$$\phi(B) = \prod_{j=1}^{p} (1 - \lambda_j B).$$

Thus $\phi(B)X_t = \varepsilon_t$ can be written as

$$\phi(B)X_t = (1-\lambda_p B)(1-\lambda_{p-1}B)\ldots(1-\lambda_1 B)X_t = \varepsilon_t.$$

Using the above representation X_t can be simulated using a recursion. For simplicity we assume $p = 2$ and $\phi(B)X_t = (1-\lambda_2 B)(1-\lambda_1 B)X_t = \varepsilon_t$. First define the AR(1) model

$$(1-\lambda_1 B)Y_{1,t} = \varepsilon_t \Rightarrow Y_{1,t} = (1-\lambda_1 B)^{-1}\varepsilon_t.$$

This gives

$$(1-\lambda_2 B)X_t = (1-\lambda_1 B)^{-1}\varepsilon_t = Y_{1,t}.$$

Thus we first simulate $\{Y_{1,t}\}_t$ using the above AR(1) method described above. We treat $\{Y_{1,t}\}_t$ as the $i$nnovations, and then simulate

$$(1-\lambda_2 B)X_t = Y_{1,t},$$

using the AR(1) method described above, but treating $\{Y_{1,t}\}_t$ as the innovations. This method can easily be generalized for any AR(p) model (with real roots). Below we describe how to do the same but when the roots are complex

Simulating an AR(2) with complex roots Suppose that X_t has a causal AR(2) representation. The roots can be complex, but since X_t is real, the roots must be conjugates ($\lambda_1 = r\exp(i\theta)$ and $\lambda_2 = r\exp(-i\theta)$). This means X_t satisfies the representation

$$(1-2r\cos(\theta)B + r^2B^2)X_t = \varepsilon_t$$

where $|r| < 1$. Now by using the same method described for simulating an AR(1), we can simulate an AR(2) model with complex roots.

In summary, by using the method for simulating AR(1) and AR(2) models we can simulate any AR(p) model with both real and complex roots.

Simulating from a fully non-causal AR model

Suppose that $\{X_t\}$ is an AR(p) model with characteristic function $\phi(B)$, whose roots lie inside the unit circle (fully non-causal). Then we can simulate X_t using the backward recursion

$$X_{t-p} = \phi_p^{-1}\left[-\phi_{p-1}X_{t-p+1} - \ldots - \phi_1 X_{t-1} + X_t\right] - \phi_p^{-1}\varepsilon_t. \tag{4.13}$$

Simulating from a non-Gaussian non-causal AR model

We now describe a method for simulating AR(p) models whose roots are both inside and outside the unit circle. The innovations should be non-Gaussian, as it makes no sense to simulate a non-causal Gaussian model and it is impossible to distinguish it from a corresponding causal Gaussian model. The method described below was suggested by former TAMU PhD student Furlong Li.

Worked example To simplify the description consider the AR(2) model where $\phi(B) = (1-\lambda_1 B)(1-\mu_1 B)$ with $|\lambda_1| < 1$ (outside unit circle) and $|\mu_1| > 1$ (inside the unit circle). Then

$$(1-\lambda_1 B)(1-\mu_1 B)X_t = \varepsilon_t.$$

Define the non-causal AR(1) model

$$(1-\mu_1 B)Y_{1,t} = \varepsilon_t.$$

And simulate $\{Y_{1,t}\}$ using a backward recursion. Then treat $\{Y_{1,t}\}$ as the innovations and simulate the causal AR(1)

$$(1-\mu_1 B)X_t = Y_{1,t}$$

using a forward recursion. This gives an AR(2) model whose roots lie inside and outside the unit circle. The same method can be generalized to any non-causal AR(p) model.

Exercise 4.7 *In the following simulations, use non-Gaussian innovations.*

(i) Simulate a stationary AR(4) process with characteristic function

$$\phi(z) = \left[1 - 0.8\exp(i\frac{2\pi}{13})z\right]\left[1 - 0.8\exp(-i\frac{2\pi}{13})z\right]\left[1 - 1.5\exp(i\frac{2\pi}{5})z\right]\left[1 - 1.5\exp(-i\frac{2\pi}{5})z\right].$$

(ii) Simulate a <u>stationary</u> AR(4) process with characteristic function

$$\phi(z) = \left[1 - 0.8\exp(i\frac{2\pi}{13})z\right]\left[1 - 0.8\exp(-i\frac{2\pi}{13})z\right]\left[1 - \frac{2}{3}\exp(i\frac{2\pi}{5})z\right]\left[1 - \frac{2}{3}\exp(-i\frac{2\pi}{5})z\right].$$

Do you observe any differences between these realisations?

`R` functions

Shumway and Stoffer (2006) and David Stoffer's website gives a comprehensive introduction to time series `R`-functions.

The function `arima.sim` simulates from a Gaussian ARIMA process. For example, `arima.sim(list(order=c(2,0,0), ar = c(1.5, -0.75)), n=150)` simulates from the AR(2) model $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$, where the innovations are Gaussian.

4.5 The ARMA model

Up to now, we have focussed on the autoregressive model. The MA(q) in many respects is a much simpler model to understand. In this case the time series is a weighted sum of independent latent variables

$$X_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\theta_{t-q} = \varepsilon_t + \sum_{j=1}^{q}\theta_j\varepsilon_{t-j}. \quad (4.14)$$

We observe that X_t is independent of any X_{t-j} where $|j| \geq q+1$. On the contrast, for an AR(p) model, there is dependence between X_t and all the time series at all other time points (we have shown above that if the AR(p) is causal, then it can be written as an MA(∞) thus the dependency at all lags). There are advantages and disadvantages of using either model. The MA(q) is independent after q lags (which may be not be viewed as realistic). But for many data sets simply fitting an AR(p) model to the data and using a model selection criterion (such as AIC), may lead to the selection of a large order p. This means the estimation of many parameters for a relatively small data sets. The AR(p) may not be parsimonious. The large order is usually chosen when the correlations tend to decay slowly and/or the autcorrelations structure is quite complex (not just monotonically decaying). However, a model involving 10-15 unknown parameters is not particularly parsimonious and more parsimonious models which can model the same behaviour would be useful.

A very useful generalisation which can be more flexible (and parsimonious) is the ARMA(p, q) model, in this case X_t has the representation

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}.$$

Definition 4.5.1 (Summary of AR, ARMA and MA models) *(i) The autoregressive AR(p) model: $\{X_t\}$ satisfies*

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t. \tag{4.15}$$

Observe we can write it as $\phi(B)X_t = \varepsilon_t$

(ii) The moving average $MA(q)$ model: $\{X_t\}$ satisfies

$$X_t = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}. \tag{4.16}$$

Observe we can write $X_t = \theta(B)\varepsilon_t$

(iii) The autoregressive moving average $ARMA(p, q)$ model: $\{X_t\}$ satisfies

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}. \tag{4.17}$$

We observe that we can write X_t as $\phi(B)X_t = \theta(B)\varepsilon_t$.

We now state some useful definitions.

Definition 4.5.2 (Causal and invertible) *Consider the $ARMA(p, q)$ model defined by*

$$X_t + \sum_{j=1}^{p} \psi_j X_{t-j} = \sum_{i=1}^{q} \theta_i \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and constant variance.

(i) An ARMA process is said to be causal if it has the representation

$$X_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}.$$

(ii) *An ARMA(p,q) process $X_t + \sum_{j=1}^{p}\psi_j X_{t-j} = \sum_{i=1}^{q}\theta_i\varepsilon_t$ (where $\{\varepsilon_t\}$ are uncorrelated random variables with mean zero and constant variance) is said to be invertible if it has the representation*

$$X_t = \sum_{j=1}^{\infty} a_j X_{t-j} + \varepsilon_t.$$

We have already given conditions underwhich an AR(p) model (and consequently) and ARMA(p,q) model is causal. We now look at when an MA(q) model is invertible (this allows us to write it as an AR(∞) process).

A worked example Consider the MA(1) process

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1},$$

where $\{\varepsilon_t\}$ are iid random variables. Our aim is understand when X_t can have an AR(∞) representation. We do this using the backshift notation. Recall $B\varepsilon_t = \varepsilon_{t-1}$ substituting this into the MA(1) model above gives

$$X_t = (1+\theta B)\varepsilon_t.$$

Thus at least formally

$$\varepsilon_t = (1+\theta B)^{-1}X_t.$$

We recall that the following equality holds

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

when $|\theta B| < 1$. Therefore if $|\theta| < 1$, then

$$\varepsilon_t = (1+B\theta)^{-1}X_t = \sum_{j=0}^{\infty}(-\theta)^j B^j X_t = \sum_{j=0}^{\infty}(-\theta)^j X_{t-j}.$$

Rearranging the above gives the AR(∞) representation

$$X_t = \sum_{j=1}^{\infty} (-\theta)^j X_{t-j} + \varepsilon_t,$$

but observe this representation only holds if $|\theta| < 1$.

Conditions for invertibility of an MA(q) The MA(q) process can be written as

$$X_t = \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t.$$

It will have an AR(∞) representation if the roots of the polynomial $\theta(z) = 1 + \sum_{j=1}^{q} \theta_j z^j$ lie outside the unit circle. Then we can write $(1 + \sum_{j=1}^{q} \theta_j z)^{-1} = \sum_{j=0}^{\infty} \phi_j z^j$ (i.e. all the roots are greater than one in absolute) and we have

$$X_t = \sum_{j=1}^{\infty} a_j X_{t-j} + \varepsilon_t.$$

Causal and invertible solutions are useful in both estimation and forecasting (predicting the future based on the current and past).

Below we give conditions for the ARMA to have a causal solution and also be invertible. We also show that the coefficients of the MA(∞) representation of X_t will decay exponentially.

Lemma 4.5.1 *Let us suppose X_t is an ARMA(p, q) process with representation given in Definition 4.5.1.*

(i) If the roots of the polynomial $\phi(z)$ lie outside the unit circle, and are greater than $(1+\delta)$ (for some $\delta > 0$), then X_t almost surely has the solution

$$X_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}, \tag{4.18}$$

where $\sum_j |b_j| < \infty$ (we note that really $b_j = b_j(\phi, \theta)$ since its a function of $\{\phi_i\}$ and $\{\theta_i\}$). Moreover for all j,

$$|b_j| \leq K\rho^j \tag{4.19}$$

for some finite constant K and $1/(1+\delta) < \rho < 1$.

(ii) If the roots of $\phi(z)$ lie both inside or outside the unit circle and are larger than $(1+\delta)$ or less than $(1+\delta)^{-1}$ for some $\delta > 0$, then we have

$$X_t = \sum_{j=-\infty}^{\infty} b_j \varepsilon_{t-j}, \tag{4.20}$$

(a vector $AR(1)$ is not possible), where

$$|a_j| \leq K\rho^{|j|} \tag{4.21}$$

for some finite constant K and $1/(1+\delta) < \rho < 1$.

(iii) If the absolute value of the roots of $\theta(z) = 1 + \sum_{j=1}^{q} \theta_j z^j$ are greater than $(1+\delta)$, then (4.17) can be written as

$$X_t = \sum_{j=1}^{\infty} a_j X_{t-j} + \varepsilon_t. \tag{4.22}$$

where

$$|a_j| \leq K\rho^{j} \tag{4.23}$$

for some finite constant K and $1/(1+\delta) < \rho < 1$.

To compare the behaviour or an AR and ARMA models we simulate from and AR(3) and and ARMA$(3,2)$ where both models have the same autoregressive parameters. We simulate from the AR(3) model (two complex roots, one real root)

$$(1 - 2\cdot 0.8\cos(\pi/3)B + 0.8^2B^2)(1-0.6B)X_t = \varepsilon_t$$

and the ARMA$(3,2)$ model

$$(1 - 2\cdot 0.8\cos(\pi/3)B + 0.8^2B^2)(1-0.6B)X_t = (1+0.5B-0.5B^2)\varepsilon_t$$

The realisations and corresponding periodogram are given in Figures 4.11 and 4.12. Observe that the AR(3) model has one real root $\lambda = 0.6$, this gives rise to the perceived curve in Figure 4.11

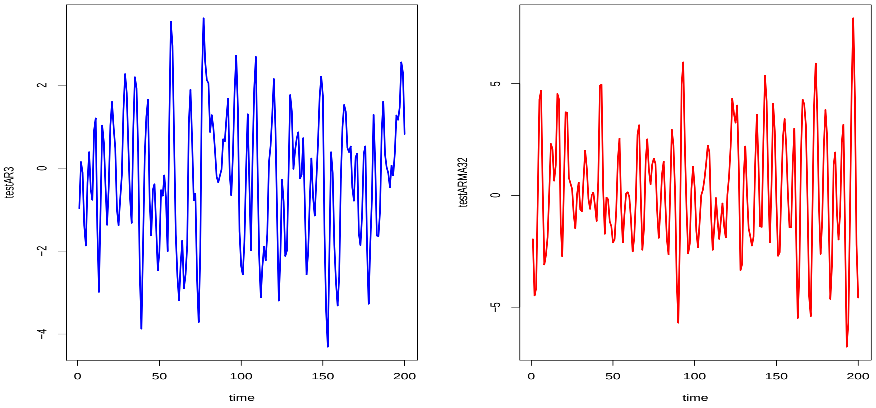


Figure 4.11: Realisation from Left: AR(3) and Right: ARMA(3, 2)

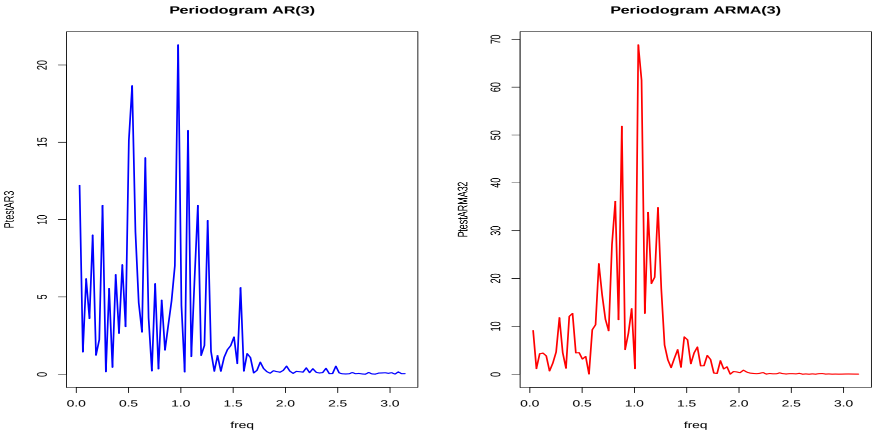


Figure 4.12: Periodogram from realisation from Left: AR(3) and Right: ARMA(3, 2)

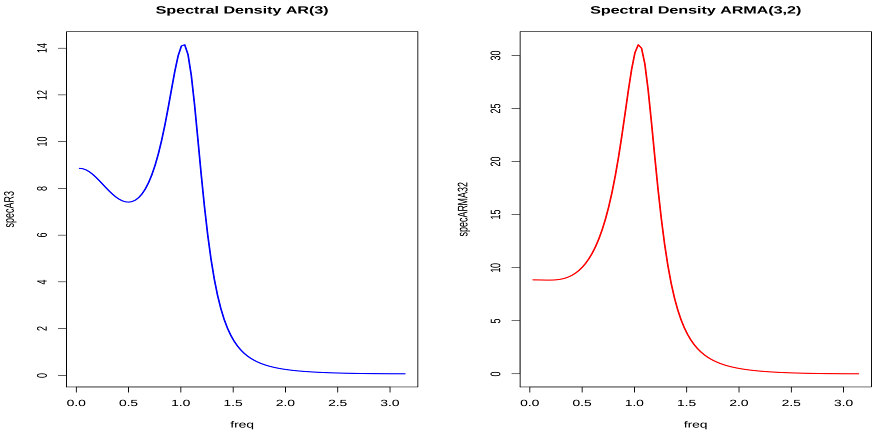


Figure 4.13: Spectral density from Left: AR(3) and Right: ARMA(3, 2)

and relatively amplitudes at low frequencies in the corresponding periodogram (in Figure 4.12). In contrast, the ARMA model has exactly the same AR part as the AR(3) model, but the MA part of this model appears to cancel out some of the low frequency information! The corresponding spectral density of the AR(3) and ARMA(3, 2) model are

$$f_{AR}(\omega) = \frac{1}{|1 - 1.6\cos\theta e^{i\omega} + 0.8^2 e^{2i\omega}|^2 |1 - 0.6e^{i\omega}|^2}$$

and

$$f_{ARMA}(\omega) = \frac{|1 + 0.5e^{i\omega} - 0.5e^{2i\omega}|^2}{|1 - 1.6\cos\theta e^{i\omega} + 0.8^2 e^{2i\omega}|^2 |1 - 0.6e^{i\omega}|^2}$$

respectively. A plot of these spectral densities is given in Figure 4.13. We observe that the periodogram maps the rough character of the spectral density. This the spectral density conveys more information than then simply being a positive function. It informs on where periodicities in the time series are most likely to lie. Studying 4.13 we observe that MA part of the ARMA spectral density appears to be dampening the low frequencies. Code for all these models is given on the course website. Simulate different models and study their behaviour.

4.6 ARFIMA models

We have shown in Lemma 4.5.1 that the coefficients of an ARMA processes which admit a stationary solution decay geometrically. This means that they are unable to model "persistant" behaviour between random variables which are separately relatively far in time. However, the ARIMA offers a solution on how this could be done. We recall that $(1-B)X_t = \varepsilon_t$ is a process which is nonstationary. However we can no replace $(1-B)^d$ (where d is a fraction) and see if one can obtain a compromise between persistance (long memory) and nonstatonary (in the sense of differencing). Suppose

$$(1 - B)^d X_t = \varepsilon_t.$$

If $0 \leq d \leq 1/2$ we have the expansions

$$(1 - B)^d = \sum_{j=0}^{\infty} \psi_j B^j \qquad (1 - B)^{-d} = \sum_{j=0}^{\infty} \phi_j B^j$$

where

$$\phi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \qquad \psi_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)}$$

and $\Gamma(1+k) = k\Gamma(k)$ is the Gamma function. Note that $\sum_{j=0}^{\infty}\psi_j^2 < \infty$ but $\sum_{j=0}^{\infty}\psi_j = \infty$. This means that X_t has the stationary solution

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

Noting to show that the above is true requires weaker conditions than those given in Lemma 4.2.1. It above process does not decay geometrically fast, and it can be shown that the sample covariance is such that $c(r) \sim |r|^{2d-1}$ (hence is not absolutely summable).

4.7 Unit roots, integrated and non-invertible processes

4.7.1 Unit roots

If the difference equation has a root which is one, then an (almost sure) stationary solution of the AR model does not exist. The simplest example is the 'random walk' $X_t = X_{t-1} + \varepsilon_t$ ($\phi(z) = (1-z)$). This is an example of an Autoregressive Integrated Moving Average ARIMA$(0,1,0)$ model $(1-B)X_t = \varepsilon_t$.

To see that it does not have a stationary solution, we iterate the equation n steps backwards; $X_t = \sum_{j=0}^{n}\varepsilon_{t-j} + X_{t-n}$. $S_{t,n} = \sum_{j=0}^{n}\varepsilon_{t-j}$ is the partial sum, but it is clear that the partial sum $S_{t,n}$ does not have a limit, since it is not a Cauchy sequence, ie. $|S_{t,n} - S_{t,m}|$ does not have a limit. However, given some initial value X_0, for $t > 0$ the so called "unit process" $X_t = X_{t-1} + \varepsilon_t$ is well defined. Notice that the nonstationary solution of this sequence is $X_t = X_0 + \sum_{j=1}^{t}\varepsilon_{t-j}$ which has variance $\text{var}(X_t) = \text{var}(X_0) + t$ (assuming that $\{\varepsilon_t\}$ are iid random variables with variance one and independent of X_0).

We observe that we can 'stationarize' the process by taking first differences, i.e. defining $Y_t = X_t - X_{t-1} = \varepsilon_t$.

Unit roots for higher order differences The unit process described above can be generalised to taking d differences (often denoted as an ARIMA$(0,d,0)$) where $(1-B)^dX_t = \varepsilon_t$ (by taking d-differences we can remove d-order polynomial trends). We elaborate on this below.

To stationarize the sequence we take d differences, i.e. let $Y_{t,0} = X_t$ and for $1 \leq i \leq d$ define the iteration

$$Y_{t,i} = Y_{t,i-1} - Y_{t-1,i-1}$$

and $Y_t = Y_{t,d}$ will be a stationary sequence. Note that this is equivalent to

$$Y_t = \sum_{j=0}^{d} \frac{d!}{j!(d-j)!}(-1)^j X_{t-j}.$$

The ARIMA(p,d,q) model The general ARIMA(p,d,q) is defined as $(1-B)^d\phi(B)X_t = \theta(B)\varepsilon_t$, where $\phi(B)$ and $\theta(B)$ are p and q order polynomials respectively and the roots of $\phi(B)$ lie outside the unit circle.

Another way of describing the above model is that after taking d differences (as detailed in (ii)) the resulting process is an ARMA(p,q) process (see Section 4.5 for the definition of an ARMA model).

To illustrate the difference between stationary ARMA and ARIMA processes, in Figure 4.14

Suppose $(1-B)\phi(B)X_t = \varepsilon_t$ and let $\widetilde{\phi}(B) = (1-B)\phi(B)$. Then we observe that $\widetilde{\phi}(1) = 0$. This property is useful when checking for unit root behaviour (see Section 4.9).

More exotic unit roots

The unit root process need not be restricted to the case that the characteristic polynomial associated the AR model is one. If the absolute of the root is equal to one, then a stationary solution cannot exist. Consider the AR(2) model

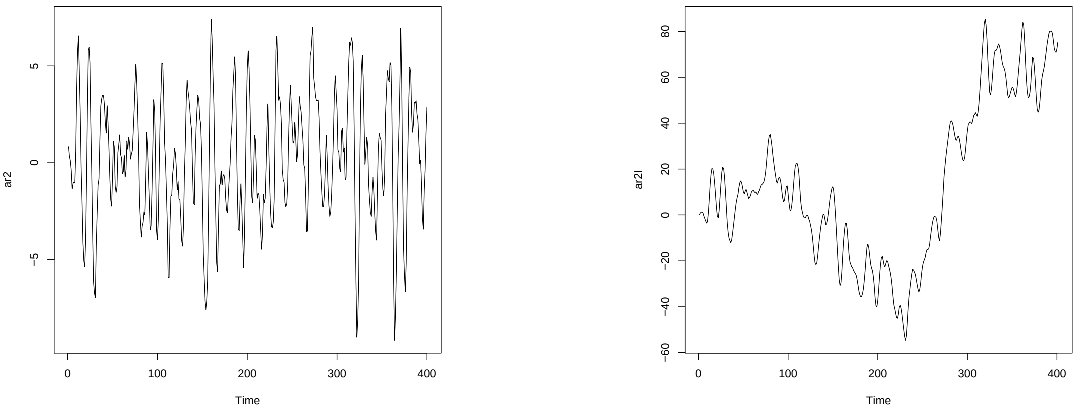
$$X_t = 2\cos\theta X_{t-1} - X_{t-2} + \varepsilon_t.$$

The associated characteristic polynomial is $\phi(B) = 1 - 2\cos(\theta)B + B^2 = (1 - e^{i\theta}B)(1 - e^{-i\theta}B)$. Thus the roots are $e^{i\theta}$ and $e^{-i\theta}$ both of which lie on the unit circle. Simulate this process.

4.7.2 Non-invertible processes

In the examples above a stationary solution does not exist. We now consider an example where the process is stationary but an autoregressive representation does not exist (this matters when we want to forecast).

Consider the MA(1) model $X_t = \varepsilon_t - \varepsilon_{t-1}$. We recall that this can be written as $X_t = \phi(B)\varepsilon_t$ where $\phi(B) = 1 - B$. From Example 4.3.1(iv) we know that $\phi(z)^{-1}$ does not exist, therefore it does not have an AR(∞) representation since $(1-B)^{-1}X_t = \varepsilon_t$ is not well defined.



(a) $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$

(b) $(1-B)Y_t = X_t$, where X_t is defined in (a)

Figure 4.14: Realisations from an AR process and its corresponding integrated process, using $N(0,1)$ innovations (generated using the same seed).

4.8 Simulating from models

4.9 Some diagnostics

Here we discuss some guidelines which allows us to discriminate between a pure autoregressive process and a pure moving average process; both with low orders. And also briefly discuss how to identify a "unit root" in the time series and whether the data has been over differenced.

4.9.1 ACF and PACF plots for checking for MA and AR behaviour

The ACF and PACF plots are the autocorrelations and partial autocorrelations estimated from the time series data (estimated assuming the time series is second order stationary). The ACF we came across is Chapter 1, the PACF we define in Chapter 6, however roughly it is the correlation between two time points after removing the linear dependence involving the observations inbetween. In `R` the functions are `acf` and `pacf`. Note that the PACF at lag zero is not given (as it does not make any sense).

The ACF and PACF of an AR(1), AR(2), MA(1) and MA(2) are given in Figures 4.15-4.18.

We observe from Figure 4.15 and 4.16 (which give the ACF of and AR(1) and AR(2) process) that there is correlation at all lags (though it reduces for large lags). However, we see from the PACF for the AR(1) has only one large coefficient at lag one and the PACF plot of the AR(2) has two large coefficients at lag one *and* two. This suggests that the ACF and PACF plot can be used to diagnose autoregressive behaviour and its order.

Similarly, we observe from Figures 4.17 and 4.18 (which give the ACF of and MA(1) and MA(2) process) that there is no real correlation in the ACF plots after lag one and two respectively, but the PACF plots are more ambigious (there seems to be correlations at several lags).

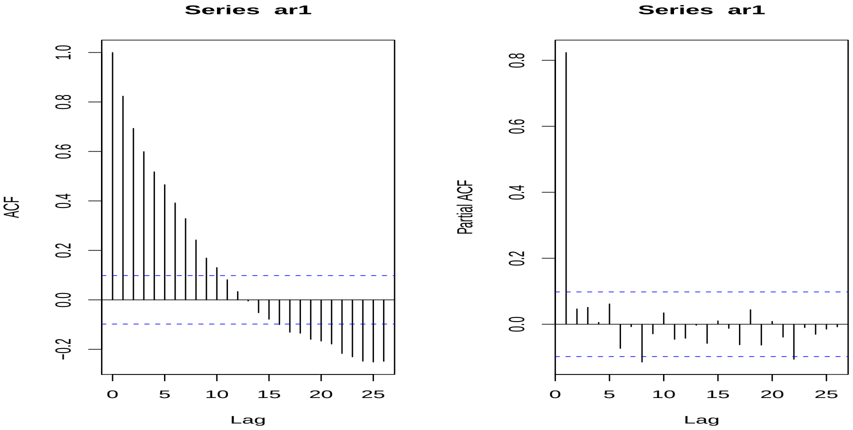


Figure 4.15: ACF and PACF plot of an AR(1), $X_t = 0.5X_{t-1} + \varepsilon_t$, $n = 400$

4.9.2 Checking for unit roots

We recall that for an AR(1) process, the unit root corresponds to $X_t = X_{t-1} + \varepsilon_t$ i.e. $\phi = 1$. Thus to check for unit root type behaviour we estimate ϕ and see how close ϕ is to one. We can formally turn this into a statistical test $H_0 : \phi = 1$ vs. $H_A : |\phi| < 1$ and there several tests for this, the most famous is the Dickey-Fuller test. Rather intriguingly, the distribution of $\widehat{\phi}$ (using the least squares estimator) does not follow a normal distribution with a $\sqrt{n}$-rate!

Extending the the unit root to the AR(p) process, the unit root corresponds to $(1-B)\phi(B)X_t = \varepsilon_t$ where $\phi(B)$ is an order $(p-1)$-polynomial (this is the same as saying $X_t - X_{t-1}$ is a stationary AR($p-1$) process). Checking for unit root is the same as checking that the sum of all the AR

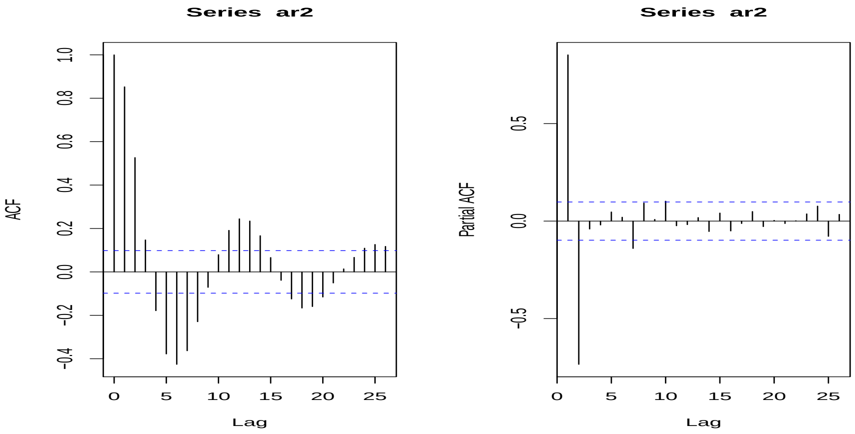


Figure 4.16: ACF and PACF plot of an AR(2), $n = 400$

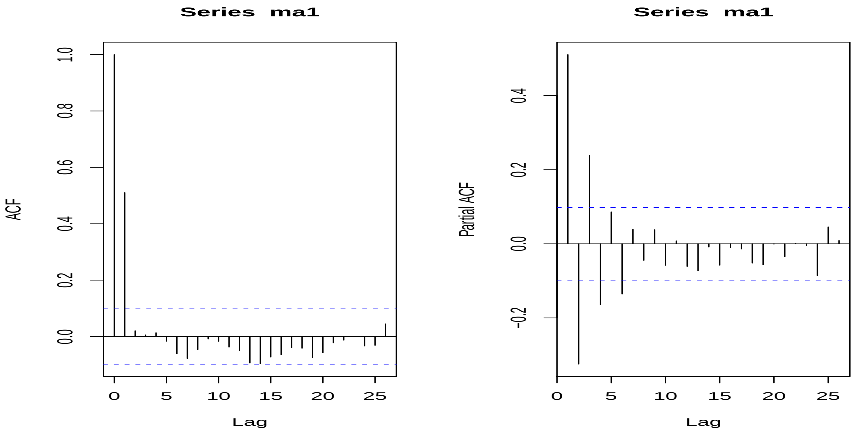


Figure 4.17: ACF and PACF plot of an MA(1), $X_t = \varepsilon_t + 0.8\varepsilon_{t-1}$, $n = 400$

coefficients is equal to one. This is easily seen by noting that $\widetilde{\phi}(1) = 0$ where $\widetilde{\phi}(B) = (1 - B)\phi(B)$ or

$$(1 - B)\phi(B)X_t = X_t - (\phi_1 - 1)X_{t-1} - (\phi_2 - \phi_1)X_{t-2} - (\phi_{p-1} - \phi_{p-2})X_{t-p+1} + \phi_{p-1}X_{t-p} = \varepsilon_t.$$

Thus we see that the sum of the AR coefficients is equal to one. Therefore to check for unit root behaviour in AR(p) processes one can see how close the sum of the estimate AR coefficients $\sum_{j=1}^{p} \widehat{\phi}_j$ is to one. Again this can be turned into a formal test.

In order to remove stochastic or deterministic trend one may difference the data. But if the data is over differenced one can induce spurious dependence in the data which is best avoided (estimation is terrible and prediction becomes a nightmare). One indicator of over differencing is

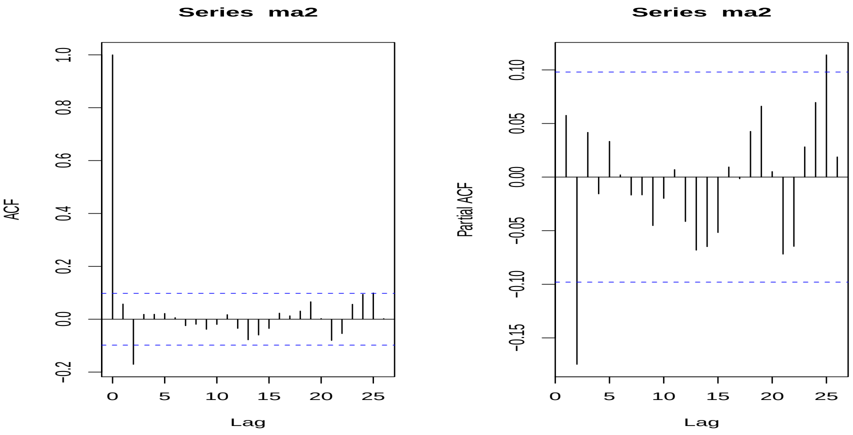


Figure 4.18: ACF and PACF plot of an MA(2), $n = 400$

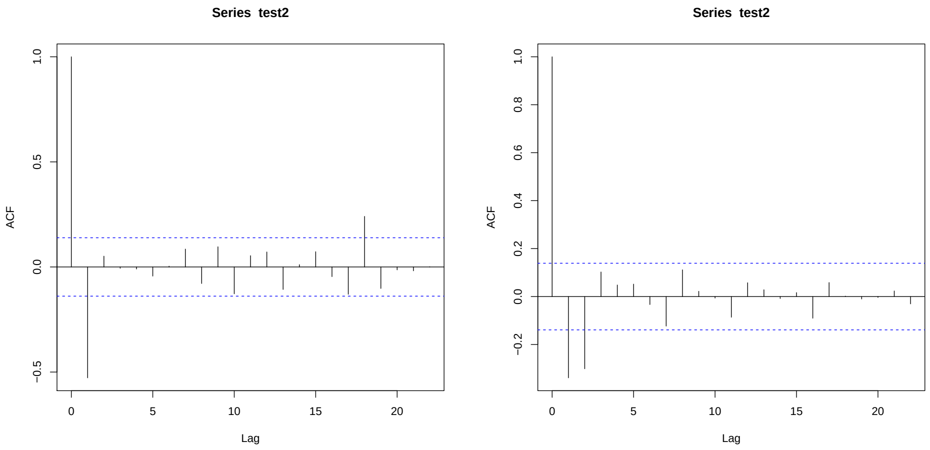


Figure 4.19: ACF of differenced data $Y_t = X_t - X_{t-1}$. Left $X_t = \varepsilon_t$, Right $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$.

the appearance of negative correlation at lag one in the data. This is illustrated in Figure 4.19, where for both data sets (difference of iid noise and differenced of an AR(2) process) we observe a large negative correlation at lag one.

4.10 Appendix

Representing an AR(p) model as a VAR(1) Let us suppose X_t is an AR(p) process, with the representation

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t.$$

For the rest of this section we will assume that the roots of the characteristic function, $\phi(z)$, lie outside the unit circle, thus the solution causal. We can rewrite the above as a Vector Autoregressive (VAR(1)) process

$$\underline{X}_t = A\underline{X}_{t-1} + \underline{\varepsilon}_t \tag{4.24}$$

where

$$\left(\begin{array}{ccccc} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 \end{array}\right), \tag{4.25}$$

$\underline{X}_t' = (X_t, \dots, X_{t-p+1})$ and $\underline{\varepsilon}_t' = (\varepsilon_t, 0, \dots, 0)$. It is straightforward to show that the eigenvalues of A are the inverse of the roots of $\phi(z)$ (since

$$\det(A - zI) = z^p - \sum_{i=1}^{p} \phi_i z^{p-i} = \underbrace{z^p\,(1 - \sum_{i=1}^{p} \phi_i z^{-i})}_{=z^p\phi(z^{-1})}),$$

thus the eigenvalues of A lie inside the unit circle. It can be shown that for any $|\lambda_{max}(A)| < \delta < 1$, there exists a constant C_δ such that $|\|A^j\|_{spec} \leq C_\delta \delta^j$ (see Appendix A). Note that result is extremely obvious if the eigenvalues are distinct (in which case the spectral decomposition can be used), in which case $|\|A^j\|_{spec} \leq C_\delta |\lambda_{max}(A)|^j$ (note that $\|A\|_{spec}$ is the spectral norm of A, which is the largest eigenvalue of the symmetric matrix AA').

We can apply the same back iterating that we did for the AR(1) to the vector AR(1). Iterating (13.4) backwards k times gives

$$\underline{X}_t = \sum_{j=0}^{k-1} A^j \underline{\varepsilon}_{t-j} + A^k \underline{X}_{t-k}.$$

Since $\|A^k \underline{X}_{t-k}\|_2 \leq \|A^k\|_{spec} \|\underline{X}_{t-k}\| \overset{\mathcal{P}}{\to} 0$ we have

$$\underline{X}_t = \sum_{j=0}^{\infty} A^j \underline{\varepsilon}_{t-j}.$$

We use the above representation to prove Lemma 4.5.1.

PROOF of Lemma 4.5.1 We first prove (i) There are several way to prove the result. The proof we consider here, uses the VAR expansion given in Section **??**; thus we avoid using the Backshift operator (however the same result can easily proved using the backshift). We write the ARMA process as a vector difference equation

$$\underline{X}_t = A\underline{X}_{t-1} + \underline{\varepsilon}_t \tag{4.26}$$

where $\underline{X}_t' = (X_t, \ldots, X_{t-p+1})$, $\underline{\varepsilon}_t' = (\varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, 0, \ldots, 0)$. Now iterating (4.26), we have

$$\underline{X}_t = \sum_{j=0}^{\infty} A^j \underline{\varepsilon}_{t-j}, \tag{4.27}$$

concentrating on the first element of the vector $\underline{X}_t$ we see that

$$X_t = \sum_{i=0}^{\infty} [A^i]_{1,1}(\varepsilon_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-i-j}).$$

Comparing (4.18) with the above it is clear that for $j > q$, $a_j = [A^j]_{1,1} + \sum_{i=1}^{q} \theta_i [A^{j-i}]_{1,1}$. Observe that the above representation is very similar to the AR(1). Indeed as we will show below the A^j behaves in much the same way as the ϕ^j in AR(1) example. As with ϕ^j, we will show that A^j converges to zero as $j \to \infty$ (because the eigenvalues of A are less than one). We now show that $|X_t| \leq K \sum_{j=1}^{\infty} \rho^j |\varepsilon_{t-j}|$ for some $0 < \rho < 1$, this will mean that $|a_j| \leq K\rho^j$. To bound $|X_t|$ we use (4.27)

$$|X_t| \leq \|\underline{X}_t\|_2 \leq \sum_{j=0}^{\infty} \|A^j\|_{spec} \|\underline{\varepsilon}_{t-j}\|_2.$$

Hence, by using Gelfand's formula (see Appendix A) we have $|\|A^j\|_{spec} \leq C_\rho \rho^j$ (for any $|\lambda_{\max}(A)| < \rho < 1$, where $\lambda_{\max}(A)$ denotes the largest maximum eigenvalue of the matrix A), which gives the corresponding bound for $|a_j|$.

To prove (ii) we use the backshift operator. This requires the power series expansion of $\frac{\theta(z)}{\phi(z)}$. If the roots of $\phi(z)$ are distinct, then it is straightforward to write $\phi(z)^{-1}$ in terms of partial fractions which uses a convergent power series for $|z| = 1$. This expansion immediately gives the the linear coefficients a_j and show that $|a_j| \leq C(1+\delta)^{-|j|}$ for some finite constant C. On the other hand, if there are multiple roots, say the roots of $\phi(z)$ are $\lambda_1, \ldots, \lambda_s$ with multiplicity $m_1, \ldots, m_s$ (where $\sum_{j=1}^{s} m_s = p$) then we need to adjust the partial fraction expansion. It can be shown that $|a_j| \leq C|j|^{\max_s |m_s|}(1+\delta)^{-|j|}$. We note that for every $(1+\delta)^{-1} < \rho < 1$, there exists a constant such that $|j|^{\max_s |m_s|}(1+\delta)^{-|j|} \leq C\rho^{|j|}$, thus we obtain the desired result.

To show (iii) we use a similar proof to (i), and omit the details. □

Corollary 4.10.1 *An ARMA process is invertible if the roots of $\theta(B)$ (the MA coefficients) lie outside the unit circle and causal if the roots of $\phi(B)$ (the AR coefficients) lie outside the unit circle.*

An AR(p) process and an MA(q) process is identifiable (meaning there is only one model associated to one solution). However, the ARMA is not necessarily identifiable. The problem arises when the characteristic polynomial of the AR and MA part of the model share common roots. A simple example is $X_t = \varepsilon_t$, this also satisfies the representation $X_t - \phi X_{t-1} = \varepsilon_t - \phi\varepsilon_{t-1}$ etc. Therefore it is not possible to identify common factors in the polynomials.

One of the main advantages of the invertibility property is in prediction and estimation. We will consider this in detail below. It is worth noting that even if an ARMA process is not invertible, one can generate a time series which has identical correlation structure but is invertible (see Section 6.4).

Chapter 5

A review of some results from multivariate analysis

5.1 Preliminaries: Euclidean space and projections

In this section we describe the notion of projections. Understanding linear predictions in terms of the geometry of projections leads to a deeper understanding of linear predictions and also algorithms for solving linear systems. We start with a short review of projections in Euclidean space.

5.1.1 Scalar/Inner products and norms

Suppose $\underline{x}_1, \ldots, \underline{x}_p \in \mathbb{R}^d$, where $p < d$. There are two important quantities associated with the space $\mathbb{R}^d$:

- The Euclidean norm: $\|\underline{x}\|_2 = \sqrt{\sum_{j=1}^{d} x_j^2}$.

  When we switch to random variables the $L2$-norm changes to the square root of the variance.

- The scalar/inner product

$$\langle \underline{x}_a, \underline{x}_b \rangle = \sum_{j=1}^{d} x_{aj} x_{bj}.$$

  If $\underline{x}_a$ and $\underline{x}_b$ are orthogonal then the angle between them is 90 degrees and $\langle \underline{x}_a, \underline{x}_b \rangle = 0$. It is clear that $\langle \underline{x}, \underline{x} \rangle = \|\underline{x}\|_2^2$.

When we switch to random variables, the inner product becomes the variance covariance. Two random variables are uncorrelated if their covariance is zero.

Let $X = \text{sp}(\underline{x}_1, \ldots, \underline{x}_p)$ denote the space spanned by the vectors $\underline{x}_1, \ldots, \underline{x}_p$. This means if $\underline{z} \in \text{sp}(\underline{x}_1, \ldots, \underline{x}_p)$, there exists coefficients $\{\alpha_j\}_{j=1}^{p}$ where $\underline{z} = \sum_{j=1}^{p} \alpha_j \underline{x}_j$.

5.1.2 Projections

Let $\underline{y} \in \mathbb{R}^d$. Our aim is to project $\underline{y}$ onto $\text{sp}(\underline{x}_1, \ldots, \underline{x}_p)$. The projection will lead to an error which is orthogonal to $\text{sp}(\underline{x}_1, \ldots, \underline{x}_p)$. The projection of $\underline{y}$ onto $\text{sp}(\underline{x}_1, \ldots, \underline{x}_p)$ is the linear combination $\underline{z} = \sum_{j=1}^{p} \alpha_j \underline{x}_j$ which minimises the Euclidean distance (least squares)

$$\left\| \underline{y} - \sum_{j=1}^{p} \alpha_j \underline{x}_j \right\|_2^2 = \langle \underline{y} - \sum_{j=1}^{p} \alpha_j \underline{x}_j, \underline{y} - \sum_{j=1}^{p} \alpha_j \underline{x}_j \rangle.$$

The coefficients $\{\alpha_j\}_{j=1}^{p}$ which minimise this difference correspond to the normal equations:

$$\langle \underline{y} - \sum_{j=1}^{p} \alpha_j \underline{x}_j, \underline{x}_\ell \rangle = \underline{y}' \underline{x}_\ell - \sum_{j=1}^{p} \alpha_j \underline{x}_j' \underline{x}_\ell = 0. \qquad 1 \leq \ell \leq p. \tag{5.1}$$

The normal equations in (5.1) can be put in matrix form

$$\begin{aligned} \underline{y}' \underline{x}_\ell - \sum_{j=1}^{p} \alpha_j \underline{x}_j' \underline{x}_\ell &= 0 \\ \Rightarrow X' X \underline{\alpha} &= X \underline{y} \end{aligned} \tag{5.2}$$

where $X' = (\underline{x}_1, \ldots, \underline{x}_p)$. This leads to the well known solution

$$\underline{\alpha} = (X'X)^{-1} X \underline{y}. \tag{5.3}$$

The above shows that the best linear predictors should be such that the error $\underline{y} - \sum_{j=1}^{p} \alpha_j \underline{x}_j$ and $\underline{x}_\ell$ are orthogonal (90 degrees). Let $X = \text{sp}(\underline{x}_1, \ldots, \underline{x}_p)$, to simplify notation we often use the notation $P_X(\underline{y})$ to denote the projection of $\underline{y}$ onto X. For example, $P_X(\underline{y}) = \sum_{j=1}^{p} \alpha_j \underline{x}_j$, where $\langle \underline{y} - P_X(\underline{y}), \underline{x}_\ell \rangle = 0$ for all $1 \leq \ell \leq p$. We will often use this notation to simplify the exposition below.

Since the projection error $\underline{y} - P_X(\underline{y})$ contains no linear information on X, then

- All information on the Inner product between $\underline{y}$ and $\underline{x}_\ell$ is contained in its projection:

$$\langle \underline{y}, \underline{x}_\ell \rangle = \underline{y}' \underline{x}_\ell = \langle P_X(\underline{y}), \underline{x}_\ell \rangle \qquad 1 \leq \ell \leq p$$

- Euclidean distance of projection error:

$$\begin{aligned} & \langle \underline{y} - P_X(\underline{y}), \underline{y} \rangle \\ = \; & \langle \underline{y} - P_X(\underline{y}), \underline{y} - P_X(\underline{y}) + P_X(\underline{y}) \rangle \\ = \; & \langle \underline{y} - P_X(\underline{y}), \underline{y} - P_X(\underline{y}) \rangle + \underbrace{\langle \underline{y} - P_X(\underline{y}), P_X(\underline{y}) \rangle}_{=0} = \left\| \underline{y} - P_X(\underline{y}) \right\|_2^2. \end{aligned}$$

5.1.3 Orthogonal vectors

We now consider the simple, but important case that the vectors $\{\underline{x}_j\}_{j=1}^p$ are orthogonal. In this case, evaluation of the coefficients $\underline{\alpha}' = (\alpha_1, \ldots, \alpha_p)$ is simple. From (5.4) we recall that

$$\underline{\alpha} = (X'X)^{-1} X \underline{y}. \tag{5.4}$$

If $\{\underline{x}_j\}_{j=1}^p$ are orthogonal, then $X'X$ is a diagonal matrix where

$$(X'X) = \mathrm{diag}\left(\underline{x}_1' \underline{x}_1, \ldots, \underline{x}_p' \underline{x}_p \right).$$

Since

$$(X\underline{y})_i = \sum_{j=1}^{d} x_{i,j} y_j.$$

This gives the very simple, entry wise solution for α_j

$$\alpha_j = \frac{\sum_{j=1}^d x_{i,j} y_j}{\sum_{j=1}^d x_{ij}^2}.$$

5.1.4 Projecting in multiple stages

Suppose that $\underline{x}_1, \ldots, \underline{x}_p, \underline{x}_{p+1} \in \mathbb{R}^d$. Let $X_p = \mathrm{sp}(\underline{x}_1, \ldots, \underline{x}_p)$ and $X_{p+1} = \mathrm{sp}(\underline{x}_1, \ldots, \underline{x}_{p+1})$. Observe that X_p is a subset of X_{p+1}. With a little thought it is clear that $X_{p+1} = \mathrm{sp}(X_p, \underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1}))$. In other words, $\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1})$ is the additional information in $\underline{x}_{p+1}$ that is not contained in X_p.

If $\underline{x}_{p+1} \in X_p$, then $\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1}) = 0$.

Let $\underline{y} \in \mathbb{R}^d$. Our aim is to project $\underline{y}$ onto X_{p+1}, but we do it in stages. By first projecting onto X_p, then onto X_{p+1}. Since $\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1})$ is orthogonal to X_p (this is by the very definition of $P_{X_p}(\underline{x}_{p+1})$) we can write

$$\begin{aligned}\underline{y} &= P_{X_p}(\underline{y}) + P_{\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1})}(\underline{y}) + \varepsilon \\ &= P_{X_p}(\underline{y}) + \alpha(\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1})) + \varepsilon.\end{aligned}$$

The coefficient α can deduced by minimising the Euclidean distance of the above;

$$\left\| \underline{y} - P_{X_p}(\underline{y}) - \alpha(\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1})) \right\|_2^2 .$$

Differentiating with respect to α leads to the normal equation

$$\begin{aligned} & \langle \underline{y} - P_{X_p}(\underline{y}) - \alpha(\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1})), (\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1})) \rangle = 0 \\ = \ & \langle \underline{y} - \alpha(\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1})), (\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1})) \rangle = 0,\end{aligned}$$

where the last line is because $P_{X_p}(\underline{x}_{p+1}))$ is orthogonal to $(\underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1}))$. Thus solving the above gives

$$\alpha = \frac{\langle \underline{y}, \underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1}) \rangle}{\| \underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1}) \|_2^2}.$$

Therefore we can write $\underline{y}$ as

$$\underline{y} = \left[P_{X_p}(\underline{y}) - \alpha P_{X_p}(\underline{x}_{p+1})) \right] + \alpha \underline{x}_{p+1} + \varepsilon. \tag{5.5}$$

If $\alpha = 0$, then $\underline{x}_{p+1}$ does not contain any additional information of $\underline{y}$ over what is already in X_p.

The above may seem a little heavy. But with a few sketches using $\mathbb{R}^3$ as an example will make the derivations obvious. Once you are comfortable with projections in Euclidean space, the same ideas transfer to projections of random variables where the innerproduct in the space is the covariances (and not the scalar product).

5.1.5 Spaces of random variables

The set-up described above can be generalized to any general vector space. Our focus will be on spaces of random variables. We assume the random variables in the appropriate probability space. We then define the (Hilbert) space of random variables

$$H = \{X; \text{X is a (real) random variables where } \mathrm{var}(X) < \infty\}.$$

This looks complicated, but in many ways it is analogous to Euclidean space. There are a few additional complications (such as showing the space is complete, which we ignore). In order to define a projection in this space project, we need to define the corresponding innerproduct and norm for this space. Suppose $X, Y \in H$, then the inner-product is the covariance

$$\langle X, Y\rangle = \mathrm{cov}(X, Y).$$

The norm is clearly the variance

$$\|X\|_2^2 = \langle X, X\rangle = \mathrm{cov}(X, X).$$

Most properties that apply to Euclidean space also apply to H. Suppose that $X_1, \ldots, X_n$ are random variables in H. We define the subspace $\mathrm{sp}(X_1, \ldots, X_n)$

$$\mathrm{sp}(X_1, \ldots, X_n) = \left\{Y; \text{where } Y = \sum_{j=1}^{p} a_j X_j\right\},$$

i.e. all all random variables $Z \in H$ which can be expressed as a linear combination of $\{X_j\}_{j=1}^{n}$. Now just as in Euclidean space you can project any $\underline{y} \in \mathbb{R}^d$ onto the subspace spanned by the vectors $\underline{x}_1, \ldots, \underline{x}_p$, we can project $Y \in H$ onto $X = \mathrm{sp}(X_1, \ldots, X_n)$. The projection is such that

$$P_X(Y) = \sum_{j=1}^{p} \alpha_j X_j,$$

where the $\underline{\alpha}' = (\alpha_1, \ldots, \alpha_p)$ are such that

$$\langle X_\ell, Y - \sum_{j=1}^{p} \alpha_j X_j\rangle = \mathrm{cov}(X_\ell, Y - \sum_{j=1}^{p} \alpha_j X_j) = 0 \qquad 1 \leq j \leq p.$$

Using the above we can show that $\underline{\alpha}$ satisfies

$$\alpha = [\mathrm{var}(\underline{X})]^{-1}\mathrm{cov}(\underline{X}, Y).$$

where $\underline{Y} = (X_1, \ldots, X_p)'$ (out of slopiness we will often use say we project onto $\underline{Y}$ rather than project onto the space spanned by $\underline{Y}$ which is $\mathrm{sp}(X_1, \ldots, X_n)$).

The properties described in Section 5.1.2 apply to H too:

- Inner product between Y and X_ℓ is contained in the projection:

$$\langle Y, X_\ell \rangle = \mathrm{cov}(Y, X_\ell) = \mathrm{cov}\left(P_X(Y), X_\ell\right) \qquad 1 \leq \ell \leq p. \tag{5.6}$$

- The projection error

$$\mathrm{cov}(Y - P_X(Y), Y) = \mathrm{var}[Y - P_X(Y)].$$

This is rather formal. We now connect this to results from multivariate analysis.

5.2 Linear prediction

Suppose $(Y, \mathbf{X})$, where $\mathbf{X} = (X_1, \ldots, X_p)$ is a random vector. The best linear predictor of Y given $\mathbf{X}$ is given by

$$\widehat{Y} = \sum_{j=1}^{p} \beta_j X_j$$

where $\boldsymbol{\beta} = \Sigma_{XX}^{-1}\Sigma_{XY}$, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\Sigma_{XX} = \mathrm{var}(\mathbf{X})$, $\Sigma_{XY} = \mathrm{cov}[\mathbf{X}, Y]$. The corresonding mean squared error is

$$\mathrm{E}\left(Y - \sum_{j=1}^{p} \beta_j X_j\right)^2 = \mathrm{E}(Y^2) - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.$$

Reason To understand why the above is true, we need to find the θ which minimises

$$\mathrm{E}\left(Y - \sum_{j=1}^{p} \theta_j X_j\right)^2,$$

we assume that X_j has zero mean. Differentiating the above wrt θ_i leads to the normal equations

$$-2\left(\mathrm{E}\left(YX_i\right)-\sum_{j=1}^{p}\theta_j\mathrm{E}\left(X_jX_i\right)\right)\qquad i=1,\ldots,p.$$

Equating to zero (since we want to find the θ_i which minimises the above) is

$$\underbrace{\mathrm{E}\left(YX_i\right)}_{=\mathrm{cov}(Y,X_i)}-\sum_{j=1}^{p}\theta_j\underbrace{\mathrm{E}\left(X_jX_i\right)}_{=\mathrm{cov}(X_i,X_j)}=0\qquad i=1,\ldots,p.$$

Writing the above as a matrix equation gives the solution

$$\underline{\beta}=\mathrm{var}\left(\mathbf{X}\right)^{-1}\mathrm{cov}\left(Y,\mathbf{X}\right)=\Sigma_{XX}^{-1}\Sigma_{XY}.$$

Substituting the above into the mean squared error gives

$$\mathrm{E}\left(Y-\sum_{j=1}^{p}\beta_jX_j\right)^2 \;=\; \mathrm{E}(Y^2)-2\mathrm{E}(Y\widehat{Y})+\mathrm{E}(\widehat{Y}^2).$$

Using that

$$Y=\widehat{Y}+e$$

where e is uncorrelated with $\{X_j\}$, thus it is uncorrelated with $\widehat{Y}$. This means $\mathrm{E}[Y\widehat{Y}]=\mathrm{E}[\widehat{Y}^2]$. Therefore

$$\begin{aligned}\mathrm{E}\left(Y-\sum_{j=1}^{p}\beta_jX_j\right)^2 &= \mathrm{E}(Y^2)-\mathrm{E}(\widehat{Y}^2)=\mathrm{E}(Y^2)-\underline{\beta}'\mathrm{var}(\mathbf{X})\underline{\beta}\\ &= \mathrm{E}(Y^2)-\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.\end{aligned}$$

5.3 Partial correlation

Suppose $\boldsymbol{X}=(X_1,\ldots,X_d)'$ is a zero mean random vector (we impose the zero mean condition to simplify notation but it's not necessary). The partial correlation is the covariance between X_i and X_j, conditioned on the other elements in the vector. In other words, the covariance between the residuals of X_i and X_j after removing their linear dependence on $\boldsymbol{X}_{-(ij)}$ (the vector not containing X_i and X_j) and the residual of X_j conditioned on $\boldsymbol{X}_{-(ij)}$. To obtain an expression for this correlation we simplify notation and let $X = X_i$, $Z = X_j$ and $Y = \boldsymbol{X}_{-(ij)}$

The notion of partial correlation can also easily be understood through projections and linear prediction (though there are other equivalent derivations). We describe this below. Let $P_Y(X)$ denote the projection of the random variable X onto the space spanned by Y. I.e. $P_Y(X)$ minimises the MSE $\mathrm{E}[X - \underline{\alpha}'Y]^2$. The partial correlation between X and Z given Y is

$$\rho_{X,Z|Y} = \frac{\mathrm{cov}(X - P_Y(X), Z - P_Y(Z))}{\sqrt{\mathrm{var}(X - P_Y(X))\mathrm{var}(Z - P_Y(Z))}}.$$

By using the results in the previous section we have

$$P_Y(X) \quad = \quad \underline{\alpha}'_{X,Y}\underline{Y} \text{ and } P_Y(Z) = \underline{\alpha}'_{Z,Y}\underline{Y}$$

where

$$\underline{\alpha}_{X,Y} = [\mathrm{var}(Y)]^{-1}\mathrm{cov}(X,Y) \text{ and } \underline{\alpha}_{Z,Y} = [\mathrm{var}(Y)]^{-1}\mathrm{cov}(Z,Y). \tag{5.7}$$

Using (5.7) we can write each of the terms in $\rho_{X,Z|Y}$ in terms of the elements of the variance matrix: i.e.

$$\begin{aligned}
\mathrm{cov}(X - P_Y(X), Z - P_Y(Z)) &= \mathrm{cov}(X,Z) - \mathrm{cov}(X,Y)'[\mathrm{var}(Y)]^{-1}\mathrm{cov}(Z,Y) \\
\mathrm{var}(X - P_Y(X)) &= \mathrm{var}(X) - \mathrm{cov}(X,Y)'[\mathrm{var}(Y)]^{-1}\mathrm{cov}(X,Y) \\
\mathrm{var}(Z - P_Y(Z)) &= \mathrm{var}(Z) - \mathrm{cov}(Z,Y)'[\mathrm{var}(Y)]^{-1}\mathrm{cov}(Z,Y).
\end{aligned}$$

<u>Relating partial correlation and the regression cofficients</u> We show how the above is related to the coefficients in linear regression. Using the two-stage projection scheme described in (5.5), but switching from Euclidean space (and scalar products) to random variables and covariances we can write

$$\begin{aligned}
X &= P_Y(X) + \beta_{Z\to X}(Z - P_Y(Z)) + \varepsilon_X \\
\text{and } Z &= P_Y(Z) + \beta_{X\to Z}(X - P_Y(X)) + \varepsilon_Z,
\end{aligned} \tag{5.8}$$

where

$$\beta_{Z \to X} = \frac{\mathrm{cov}(X, Z - P_Y(Z))}{\mathrm{var}(Z - P_Y(Z))} \text{ and } \beta_{X \to Z} = \frac{\mathrm{cov}(Z, X - P_Y(X))}{\mathrm{var}(X - P_Y(X))}.$$

Since $Z - P_Y(Z)$ is orthogonal to Y (and thus $\mathrm{cov}(Z - P_Y(Z), P_Y(X)) = 0$) we have

$$\mathrm{cov}(X, Z - P_Y(Z)) = \mathrm{cov}(X - P_Y(X), Z - P_Y(Z)).$$

This is the partial covariance (as it is the covariance of the residials after projecting onto Y). This links $\beta_{Z \to X}$ and $\beta_{X \to Z}$ to the partial covariance, since

$$\beta_{Z \to X} = \frac{\mathrm{cov}(X - P_Y(X), Z - P_Y(Z))}{\mathrm{var}(Z - P_Y(Z))} \text{ and } \beta_{X \to Z} = \frac{\mathrm{cov}(Z - P_Y(Z), X - P_Y(X))}{\mathrm{var}(X - P_Y(X))}.$$

To connect the regression coefficients to the partial correlations we rewrite we rewrite the partial covariance in terms of the partial correlation:

$$\mathrm{cov}(X - P_Y(X), Z - P_Y(Z)) = \rho_{X,Z|Y} \sqrt{\mathrm{var}(X - P_Y(X))\mathrm{var}(Z - P_Y(Z))}.$$

Substituting the expression for $\mathrm{cov}(X - P_Y(X), Z - P_Y(Z))$ into the expression for $\beta_{Z \to X}$ and $\beta_{X \to Z}$ gives

$$\beta_{Z \to X} = \rho_{X,Z|Y} \sqrt{\frac{\mathrm{var}(X - P_Y(X))}{\mathrm{var}(Z - P_Y(Z))}} \text{ and } \beta_{X \to Z} = \rho_{X,Z|Y} \sqrt{\frac{\mathrm{var}(Z - P_Y(Z))}{\mathrm{var}(X - P_Y(X))}}. \tag{5.9}$$

This leads to the linear regressions

$$\begin{aligned} X &= \underbrace{(P_Y(X) - \beta_{Z \to X} P_Y(Z))}_{\text{in terms of } Y} + \underbrace{\beta_{Z \to X} Z}_{\text{in terms of } Z} + \varepsilon_X \\ Z &= \underbrace{(P_Y(Z) - \beta_{X \to Z} P_Y(X))}_{\text{in terms of } Z} + \underbrace{\beta_{X \to Z} X}_{\text{in terms of } X} + \varepsilon_Z. \end{aligned}$$

For below, keep in mind that $\mathrm{var}[\varepsilon_X] = \mathrm{var}[X - P_{Y,Z}(X)]$ and $\mathrm{var}[\varepsilon_Z] = \mathrm{var}[Z - P_{Y,X}(Z)]$.

The identity in (5.9) relates the regression coefficients to the partial correlation. In particular, the partial correlation is zero if an only if the corresponding regression coefficient is zero too.

We now rewrite (5.9) in terms of $\mathrm{var}[\varepsilon_X] = \mathrm{var}[X - P_{Y,Z}(X)]$ and $\mathrm{var}[\varepsilon_Z] = \mathrm{var}[Z - P_{Y,X}(Z)]$.

This requires the following identity

$$\frac{\text{var}(X - P_{Y,Z}(X))}{\text{var}(Z - P_{Y,X}(Z))} = \frac{\text{var}(X - P_{Y}(X))}{\text{var}(Z - P_{Y}(Z))}, \tag{5.10}$$

a proof of this identity is given at the end of this section. Using this identity together with (5.9) gives

$$\beta_{Z \to X} = \rho_{X,Z|Y}\sqrt{\frac{\text{var}(\varepsilon_X)}{\text{var}(\varepsilon_Z)}} \text{ and } \beta_{X \to Z} = \rho_{X,Z|Y}\sqrt{\frac{\text{var}(\varepsilon_Z)}{\text{var}(\varepsilon_X)}} \tag{5.11}$$

and

$$\rho_{X,Z|Y} = \beta_{Z \to X}\sqrt{\frac{\text{var}(\varepsilon_Z)}{\text{var}(\varepsilon_X)}} = \beta_{X \to Z}\sqrt{\frac{\text{var}(\varepsilon_Y)}{\text{var}(\varepsilon_Z)}} \tag{5.12}$$

Proof of identity (5.10) We recall that

$$\begin{aligned} X_i &= P_{\underline{X}_{-(i,j)}}(X_i) + \beta_{ij}(X_j - P_{\underline{X}_{-(i,j)}}(X_j)) + \varepsilon_i \\ X_j &= P_{\underline{X}_{-(i,j)}}(X_j) + \beta_{ji}(X_i - P_{\underline{X}_{-(i,j)}}(X_i)) + \varepsilon_j. \end{aligned}$$

To relate $\text{var}(\varepsilon_i)$ and $\text{var}(\varepsilon_{i,-j})$ we evaluate

$$\begin{aligned} \text{var}(\varepsilon_{i,-j}) &= \text{var}(X_i - P_{\underline{X}_{-(i,j)}}(X_i)) \\ &= \text{var}[\beta_{ij}(X_j - P_{\underline{X}_{-(i,j)}}(X_j))] + \text{var}(\varepsilon_i) \\ &= \beta_{ij}^2 \text{var}[X_j - P_{\underline{X}_{-(i,j)}}(X_j)] + \text{var}(\varepsilon_i) \\ &= \frac{[\text{cov}(X_i, X_j - P_{\underline{X}_{-(i,j)}}(X_j))]^2}{\text{var}[X_j - P_{\underline{X}_{-(i,j)}}(X_j)]} + \text{var}(\varepsilon_i) \\ &= \frac{[\text{cov}(X_i - P_{\underline{X}_{-(i,j)}}(X_i), X_j - P_{\underline{X}_{-(i,j)}}(X_j))]^2}{\text{var}[X_j - P_{\underline{X}_{-(i,j)}}(X_j)]} + \text{var}(\varepsilon_i) \\ &= \frac{c_{ij}^2}{\text{var}(\varepsilon_{j,-i})} + \text{var}(\varepsilon_i). \end{aligned}$$

where $c_{ij} = \text{cov}(X_i - P_{\underline{X}_{-(i,j)}}(X_i), X_j - P_{\underline{X}_{-(i,j)}}(X_j))$. By the same argument we have

$$\begin{aligned} \text{var}(\varepsilon_{j,-i}) &= \frac{c_{ij}^2}{\text{var}(\varepsilon_{i,-j})} + \text{var}(\varepsilon_j) \\ \Rightarrow \rho_{ij}^2 &= \text{var}(\varepsilon_{j,-i})\text{var}(\varepsilon_{i,-j}) - \text{var}(\varepsilon_j)\text{var}(\varepsilon_{i,-j}). \end{aligned}$$

Putting these two equations together gives

$$\mathrm{var}(\varepsilon_{j,-i})\mathrm{var}(\varepsilon_{i,-j}) - \mathrm{var}(\varepsilon_{j})\mathrm{var}(\varepsilon_{i,-j}) = \mathrm{var}(\varepsilon_{i,-j})\mathrm{var}(\varepsilon_{j,-i}) - \mathrm{var}(\varepsilon_{i})\mathrm{var}(\varepsilon_{j,-i}).$$

This leads to the required identity

$$\frac{\mathrm{var}(\varepsilon_i)}{\mathrm{var}(\varepsilon_j)} = \frac{\mathrm{var}(\varepsilon_{i,-j})}{\mathrm{var}(\varepsilon_{j,-i})},$$

and the desired result. □

Example 5.3.1 *Define the three random vectors X_1, X_2 and X_3, where X_1 and X_2 are such that*

$$X_1 = X_3 + \varepsilon_1 \qquad X_2 = X_3 + \varepsilon_2$$

where ε_1 is independent of X_2 and X_3 and ε_2 is independent of X_1 and X_3 (and of course they are independent of each other). Then $\mathrm{cov}(X_1, X_2) = \mathrm{var}(X_3)$ *however the partial covariance between X_1 and X_2 conditioned on X_3 is zero. I.e. X_3 is driving the dependence between the models, once it is removed they are uncorrelated and, in this example, independent.*

5.4 Properties of the precision matrix

5.4.1 Summary of results

Suppose $\boldsymbol{X}' = (X_1, \ldots, X_d)$ is a zero mean random vector (we impose the zero mean condition to simplify notation but it is not necessary), where

$$\Sigma = \mathrm{var}[\boldsymbol{X}] \text{ and } \Gamma = \Sigma^{-1}.$$

Σ is called the variance matrix, Γ is called the precision matrix. Unless stated otherwise all vectors are column vectors. We summarize the main results above in the bullet points below. We then relate these quantities to the precision matrix.

- $\boldsymbol{\beta}_i' = (\beta_{i,1}, \ldots, \beta_{i,d})$ are the coefficients which minimise $\mathrm{E}[X_i - \boldsymbol{\beta}_i'\mathbf{X}_{-i}]^2$, where $\mathbf{X}_{-i}' = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$ (all elements in $\mathbf{X}$ excluding X_i).

- $\boldsymbol{\beta}'_{i,-j}$ are the coefficients which minimise $\mathrm{E}[X_i - \boldsymbol{\beta}'_{i,-j}\mathbf{X}_{-(i,j)}]^2$, where $\mathbf{X}_{-(i,j)}$ are all elements in $\mathbf{X}$ excluding X_i and X_j.

- The partial correlation between X_i and X_j is defined as

  $$\rho_{i,j} = \mathrm{cor}(\varepsilon_{i,-j}, \varepsilon_{j,-i}) = \frac{\mathrm{cov}(\varepsilon_{i,-j}, \varepsilon_{j,-i})}{\sqrt{\mathrm{var}(\varepsilon_{i,-j})\mathrm{var}(\varepsilon_{j,-i})}},$$

  where

  $$\begin{aligned} \varepsilon_{i,-j} &= X_i - \beta_{i,-j}\mathbf{X}_{-(i,j)} \\ \varepsilon_{j,-i} &= X_j - \beta_{j,-i}\mathbf{X}_{-(i,j)}. \end{aligned}$$

  It can be shown that

  $$\begin{aligned} \mathrm{cov}\,(\varepsilon_{i,-j}, \varepsilon_{j,-i}) &= \mathrm{cov}(X_i, X_j) - \mathrm{cov}(X_i, \mathbf{X}'_{-(i,j)})\mathrm{var}[\mathbf{X}_{-(i,j)}]^{-1}\mathrm{cov}(X_j, \mathbf{X}_{-(i,j)}) \\ \mathrm{var}\,(\varepsilon_{i,-j}) &= \mathrm{var}(X_i) - \mathrm{cov}(X_i, \mathbf{X}'_{-(i,j)})\mathrm{var}[\mathbf{X}_{-(i,j)}]^{-1}\mathrm{cov}(X_i, \mathbf{X}_{-(i,j)}) \\ \mathrm{var}\,(\varepsilon_{j,-i}) &= \mathrm{var}(X_j) - \mathrm{cov}(X_j, \mathbf{X}'_{-(i,j)})\mathrm{var}[\mathbf{X}_{-(i,j)}]^{-1}\mathrm{cov}(X_j, \mathbf{X}_{-(i,j)}). \end{aligned}$$

- The regression coefficients and partial correlation are related through the identity

  $$\beta_{ij} = \rho_{ij}\sqrt{\frac{\mathrm{var}(\varepsilon_i)}{\mathrm{var}(\varepsilon_j)}}. \tag{5.13}$$

Let $\Gamma_{i,j}$ denote the $(i,j)th$ entry in the precision matrix $\Gamma = \Sigma^{-1}$. Then $\Gamma_{i,j}$ satisifies the following well known properties

$$\Gamma_{ii} = \frac{1}{\mathrm{E}[X_i - \boldsymbol{\beta}'_i\mathbf{X}_{-i}]^2}.$$

For $i \neq j$ we have $\Gamma_{i,j} = -\beta_{i,j}/\mathrm{E}[X_i - \boldsymbol{\beta}'_i\mathbf{X}_{-i}]^2$ and

$$\beta_{i,j} = -\frac{\Gamma_{i,j}}{\Gamma_{ii}} \quad \text{and} \quad \rho_{i,j} = -\frac{\Gamma_{i,j}}{\sqrt{\Gamma_{ii}\Gamma_{jj}}}.$$

5.4.2 Proof of results

Regression and the precision matrix The precision matrix contains many hidden treasures. We start by showing that the entries of the precision matrix contain the regression coefficients of X_i regressed on the random vector $\mathbf{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$. We will show that the ith row of Σ^{-1} is

$$\begin{pmatrix} -\beta_{i1}/\sigma_i^2, & -\beta_{i2}/\sigma_i^2, & \ldots & 1/\sigma_i^2 & \ldots & -\beta_{id}/\sigma_i^2. \end{pmatrix}$$

where $\sigma_2^2 = \mathrm{E}[X_i - \boldsymbol{\beta}_i' \mathbf{X}_{-i}]^2$, $\sum_{j \neq i} \beta_{ij} X_{ij}$ is the best linear predictor of X_i given $\mathbf{X}_{-i}$ and the ith entry is $1/\sigma_i^2$ (notation can be simplified if set $\beta_{ii} = -1$). And equivalently the ith column of Σ^{-1} is the transpose of the vector

$$\begin{pmatrix} -\beta_{i1}/\sigma_i^2 & -\beta_{i2}/\sigma_i^2 & \ldots & -\beta_{id}/\sigma_i^2. \end{pmatrix}.$$

Though it may seem surprising at first, the result is very logical.

We recall that the coefficients $\boldsymbol{\beta}_i = (\beta_{i,1}, \ldots, \beta_{i,d})$ are the coefficients which minimise $\mathrm{E}[X_i - \boldsymbol{\beta}_i' \mathbf{X}_{-i}]^2$. This is equivalent to the derivative of the MSE being zero, this gives rise to the classical normal equations

$$\mathrm{E}[(X_i - \sum_{j \neq i} \beta_{i,j} X_j) X_\ell] = \Sigma_{i,\ell} - \sum_{j \neq i} \beta_{i,j} \Sigma_{j,\ell} = 0 \qquad 1 \leq \ell \leq j, \ell \neq i.$$

Further, since $X_i - \sum_{j \neq i} \beta_{i,j} X_j$ is orthogonal to X_j we have

$$\mathrm{E}[(X_i - \sum_{j \neq i} \beta_{i,j} X_j) X_i] = \mathrm{E}[(X_i - \sum_{j \neq i} \beta_{i,j} X_j)^2].$$

Recall that each row in the precision matrix is orthogonal to all the columns in Σ except one. We show below that this corresponds to precisely the normal equation. It is easiest seen through the the simple example of a 4 × 4 variance matrix

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{pmatrix}$$

and the corresponding regression matrix

$$\begin{pmatrix} 1 & -\beta_{12} & -\beta_{13} & -\beta_{14} \\ -\beta_{21} & 1 & -\beta_{23} & -\beta_{24} \\ -\beta_{31} & -\beta_{32} & 1 & -\beta_{34} \\ -\beta_{41} & -\beta_{42} & -\beta_{43} & 1 \end{pmatrix}.$$

We recall from the definition of β_1 that the inner product between $\mathbf{c} = (c_{11}, c_{12}, c_{13}, c_{14})$ and $\widetilde{\boldsymbol{\beta}}_1 = (1, -\beta_{12}, -\beta_{13}, -\beta_{14})$ is

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_1 \mathbf{c}_1' = \langle \widetilde{\boldsymbol{\beta}}_1, \mathbf{c}_1 \rangle &= c_{11} - \beta_{12}c_{12} - \beta_{13}c_{13} - \beta_{13}c_{13} \\ &= \mathrm{E}[(X_1 - \sum_{j=2}^{4} \beta_{1,j}X_j)X_1] = \mathrm{E}[(X_1 - \sum_{j=2}^{4} \beta_{1,j}X_j)^2]. \end{aligned}$$

Similarly

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_1 \mathbf{c}_2' = \langle \widetilde{\beta}_1, \mathbf{c}_2 \rangle &= c_{21} - \beta_{12}c_{22} - \beta_{13}c_{23} - \beta_{13}c_{23} \\ &= \mathrm{E}[(X_1 - \sum_{j=2}^{4} \beta_{1,j}X_j)X_2] = 0. \end{aligned}$$

The same is true for the other $\mathbf{c}_j$ and $\widetilde{\boldsymbol{\beta}}_j$. Based on these observations, we observe that the regression coefficients/normal equations give the orthogonal projections and

$$\begin{pmatrix} 1 & -\beta_{12} & -\beta_{13} & -\beta_{14} \\ -\beta_{21} & 1 & -\beta_{23} & -\beta_{24} \\ -\beta_{31} & -\beta_{32} & 1 & -\beta_{34} \\ -\beta_{41} & -\beta_{42} & -\beta_{43} & 1 \end{pmatrix} \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{pmatrix} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2),$$

where $\sigma_j^2 = \mathrm{E}[(X_j - \sum_{i \neq j} \beta_i X_i)^2]$. Therefore the inverse of Σ is

$$\begin{aligned}
\Sigma^{-1} &= \mathrm{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)^{-1} \begin{pmatrix} 1 & -\beta_{12} & -\beta_{13} & -\beta_{14} \\ -\beta_{21} & 1 & -\beta_{23} & -\beta_{24} \\ -\beta_{31} & -\beta_{32} & 1 & -\beta_{34} \\ -\beta_{41} & -\beta_{42} & -\beta_{43} & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1/\sigma_1^2 & -\beta_{12}/\sigma_1^2 & -\beta_{13}/\sigma_1^2 & -\beta_{14}/\sigma_1^2 \\ -\beta_{21}/\sigma_2^2 & 1/\sigma_2^2 & -\beta_{23}/\sigma_2^2 & -\beta_{24}/\sigma_2^2 \\ -\beta_{31}/\sigma_3^2 & -\beta_{32}/\sigma_3^2 & 1/\sigma_3^2 & -\beta_{34}/\sigma_3^2 \\ -\beta_{41}/\sigma_4^2 & -\beta_{42}/\sigma_4^2 & -\beta_{43}/\sigma_4^2 & 1/\sigma_4^2 \end{pmatrix}.
\end{aligned}$$

By a similar argument we have

$$\begin{aligned}
\Sigma^{-1} &= \begin{pmatrix} 1 & -\beta_{21} & -\beta_{31} & -\beta_{41} \\ -\beta_{12} & 1 & -\beta_{32} & -\beta_{42} \\ -\beta_{13} & -\beta_{23} & 1 & -\beta_{43} \\ -\beta_{14} & -\beta_{24} & -\beta_{34} & 1 \end{pmatrix} \mathrm{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)^{-1} \\
&= \begin{pmatrix} 1/\sigma_1^2 & -\beta_{21}/\sigma_2^2 & -\beta_{31}/\sigma_3^2 & -\beta_{41}/\sigma_4^2 \\ -\beta_{12}/\sigma_1^2 & 1/\sigma_2^2 & -\beta_{32}/\sigma_3^2 & -\beta_{42}/\sigma_4^2 \\ -\beta_{13}/\sigma_1^2 & -\beta_{23}/\sigma_2^2 & 1/\sigma_3^2 & -\beta_{43}/\sigma_4^2 \\ -\beta_{14}/\sigma_1^2 & -\beta_{24}/\sigma_2^2 & -\beta_{34}/\sigma_3^2 & 1/\sigma_4^2 \end{pmatrix}.
\end{aligned}$$

In summary, the normal equations give the matrix multiplication required for a diagonal matrix (which is is exactly the definition of $\Sigma\Gamma = I$, up to a change in the diagonal).

Clearly, the above proof holds for all dimensions and we have

$$\Gamma_{ii} = \frac{1}{\sigma_i^2},$$

and

$$\Gamma_{ij} = -\frac{\beta_{ij}}{\sigma_i^2} \Rightarrow \beta_{i,j} = -\frac{\Gamma_{ij}}{\Gamma_{ii}}.$$

Writing the partial correlation in terms of elements of the precision matrix By using the identity

(5.12) (and that $\beta_{ij} = \beta_{j\to i}$) we have

$$\rho_{ij} = \beta_{ij}\sqrt{\frac{\mathrm{var}[\varepsilon_j]}{\mathrm{var}[\varepsilon_i]}}. \tag{5.14}$$

We recall that $\Gamma_{ii} = \mathrm{var}(X_i - P_{X_{-i}}(X_i))^{-1}$, $\Gamma_{jj} = \mathrm{var}(X_j - P_{X_{-j}}(X_j))^{-1}$ and $\Gamma_{ij} = -\beta_{ij}\Gamma_{ii}$ gives

$$\rho_{ij} \quad = \quad -\frac{\Gamma_{ij}}{\Gamma_{ii}}\sqrt{\frac{\Gamma_{ii}}{\Gamma_{jj}}} = -\frac{\Gamma_{ij}}{\sqrt{\Gamma_{ii}\Gamma_{jj}}}.$$

The above represents the partial correlation in terms of entries of the precision matrix.

5.5 Appendix

Alternative derivations based on matrix identities

The above derivations are based on properties of normal equations and some algebraic manipulations. An alternative set of derivations is given in terms of the inversions of block matrices, specifically with the classical matrix inversions identities

$$\begin{aligned}\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} &= \begin{pmatrix} A^{-1} + A^{-1}BP_1^{-1}CA^{-1} & -A^{-1}BP_1^{-1} \\ -P_1^{-1}CA^{-1} & P_1^{-1} \end{pmatrix} \\ &= \begin{pmatrix} P_2^{-1} & -P_2^{-1}BD^{-1} \\ -D^{-1}CP_2^{-1} & D^{-1} + D^{-1}CP_2^{-1}BD^{-1} \end{pmatrix},\end{aligned} \tag{5.15}$$

where $P_1 = (D - CA^{-1}B)$ and $P_2 = (A - BD^{-1}C)$. Or using the idea of normal equations in projections.

The precision matrix and partial correlation

Let us suppose that $\boldsymbol{X} = (X_1, \ldots, X_d)$ is a zero mean random vector with variance Σ. The $(i,j)th$ element of Σ the covariance $\mathrm{cov}(X_i, X_j) = \Sigma_{ij}$. Here we consider the inverse of Σ, and what information the $(i,j)th$ of the inverse tells us about the correlation between X_i and X_j. Let Σ^{ij} denote the $(i,j)th$ element of Σ^{-1}. We will show that with appropriate standardisation, Σ^{ij} is the negative partial correlation between X_i and X_j. More precisely,

$$\frac{\Sigma^{ij}}{\sqrt{\Sigma^{ii}\Sigma^{jj}}} = -\rho_{ij}. \tag{5.16}$$

The proof uses the inverse of block matrices. To simplify the notation, we will focus on the $(1,2)th$ element of Σ and Σ^{-1} (which concerns the correlation between X_1 and X_2).

Remark 5.5.1 *Remember the reason we can always focus on the top two elements of* $\mathbf{X}$ *is because we can always use a permutation matrix to permute the* X_i *and* X_j *such that they become the top two elements. Since the inverse of the permutation matrix is simply its transpose everything still holds.*

Let $\boldsymbol{X}_{1,2} = (X_1, X_2)'$, $\boldsymbol{X}_{-(1,2)} = (X_3, \ldots, X_d)'$, $\Sigma_{-(1,2)} = \mathrm{var}(\boldsymbol{X}_{-(1,2)})$, $\underline{c}_{1,2} = \mathrm{cov}(\boldsymbol{X}_{(1,2)}, \boldsymbol{X}_{-(1,2)})$ and $\Sigma_{1,2} = \mathrm{var}(\boldsymbol{X}_{1,2})$. Using this notation it is clear that

$$\mathrm{var}(\boldsymbol{X}) = \Sigma = \begin{pmatrix} \Sigma_{1,2} & \underline{c}_{1,2} \\ \underline{c}'_{1,2} & \Sigma_{-(1,2)} \end{pmatrix}. \tag{5.17}$$

By using (5.15) we have

$$\Sigma^{-1} = \begin{pmatrix} P^{-1} & -P^{-1}\underline{c}'_{1,2}\Sigma^{-1}_{-(1,2)} \\ -\Sigma^{-1}_{-(1,2)}\underline{c}_{1,2}P^{-1} & P^{-1} + \Sigma^{-1}_{-(1,2)}\underline{c}_{1,2}P^{-1}\underline{c}'_{1,2}\Sigma^{-1}_{-(1,2)} \end{pmatrix}, \tag{5.18}$$

where $P = (\Sigma_{1,2} - \underline{c}'_{1,2}\Sigma^{-1}_{-(1,2)}\underline{c}_{1,2})$. Comparing P with (**??**), we see that P is the 2×2 variance/-covariance matrix of the residuals of $X_{(1,2)}$ conditioned on $\boldsymbol{X}_{-(1,2)}$. Thus the partial correlation between X_1 and X_2 is

$$\rho_{1,2} = \frac{P_{1,2}}{\sqrt{P_{1,1}P_{2,2}}} \tag{5.19}$$

where P_{ij} denotes the elements of the matrix P. Inverting P (since it is a two by two matrix), we see that

$$P^{-1} = \frac{1}{P_{1,1}P_{2,2} - P_{1,2}^2} \begin{pmatrix} P_{2,2} & -P_{1,2} \\ -P_{1,2} & P_{11} \end{pmatrix}. \tag{5.20}$$

Thus, by comparing (5.18) and (5.20) and by the definition of partial correlation given in (5.19) we have

$$\frac{P^{(1,2)}}{\sqrt{P^{(1,1)}P^{(2,2)}}} = -\rho_{1,2}.$$

Let Σ^{ij} denote the (i,j)th element of Σ^{-1}. Thus we have shown (5.16):

$$\rho_{ij} = -\frac{\Sigma^{ij}}{\sqrt{\Sigma^{ii}\Sigma^{jj}}}. \tag{5.21}$$

In other words, the (i,j)th element of Σ^{-1} divided by the square root of its diagonal gives negative partial correlation. Therefore, if the partial correlation between X_i and X_j given $\mathbf{X}_{ij}$ is zero, then $\Sigma^{i,j} = 0$.

The precision matrix and the coefficients in regression

The precision matrix, Σ^{-1}, contains many other hidden treasures. For example, the coefficients of Σ^{-1} convey information about the best linear predictor X_i given $\boldsymbol{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$ (all elements of $\boldsymbol{X}$ except X_i). Let

$$X_i = \sum_{j \neq i} \beta_{i,j} X_j + \varepsilon_i,$$

where $\{\beta_{i,j}\}$ are the coefficients of the best linear predictor. Then it can be shown that

$$\beta_{i,j} = -\frac{\Sigma^{ij}}{\Sigma^{ii}} \quad \text{and} \quad \Sigma^{ii} = \frac{1}{\mathrm{E}[X_i - \sum_{j \neq i} \beta_{i,j} X_j]^2}. \tag{5.22}$$

The precision matrix and the mean squared prediction error

We start with a well known expression, which expresses the prediction errors in terms of the determinant of matrices.

We recall that the prediction error is

$$\mathrm{E}[Y - \widehat{Y}]^2 = \sigma_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \tag{5.23}$$

with $\sigma_Y = \text{var}[Y]$. Let

$$\Sigma = \begin{pmatrix} \text{var}[Y] & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}. \tag{5.24}$$

We show below that the prediction error can be rewritten as

$$\text{E}[Y - \widehat{Y}]^2 = \sigma_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} = \frac{\det(\Sigma)}{\det(\Sigma_{XX})}. \tag{5.25}$$

Furthermore,

$$\left(\Sigma^{-1}\right)_{11} = \frac{1}{\sigma_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}} = \frac{1}{\text{E}[Y - \widehat{Y}]^2}. \tag{5.26}$$

Proof of (5.25) and (5.26) To prove this result we use

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(D)\det\left(A - BD^{-1}C\right). \tag{5.27}$$

Applying this to (5.27) gives

$$\begin{aligned} \det(\Sigma) &= \det(\Sigma_{XX})\left(\sigma_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right) \\ \Rightarrow \det(\Sigma) &= \det(\Sigma_{XX})\text{E}[Y - \widehat{Y}]^2, \end{aligned} \tag{5.28}$$

thus giving (5.25).

To prove (5.26) we use the following result on the inverse of block matrices

$$\begin{aligned} \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} &= \begin{pmatrix} A^{-1} + A^{-1}BP_1^{-1}CA^{-1} & -A^{-1}BP_1^{-1} \\ -P_1^{-1}CA^{-1} & P_1^{-1} \end{pmatrix} \\ &= \begin{pmatrix} P_2^{-1} & -P_2^{-1}BD^{-1} \\ -D^{-1}CP_2^{-1} & D^{-1} + D^{-1}CP_2^{-1}BD^{-1} \end{pmatrix}, \end{aligned} \tag{5.29}$$

where $P_1 = (D - CA^{-1}B)$ and $P_2 = (A - BD^{-1}C)$. This block inverse turns out to be crucial in deriving many of the interesting properties associated with the inverse of a matrix. We now show that the the inverse of the matrix Σ, Σ^{-1} (usually called the precision matrix) contains the mean squared error.

Comparing the above with (5.24) and (5.23) we see that

$$\left(\Sigma^{-1}\right)_{11} = \frac{1}{\sigma_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}} = \frac{1}{\mathrm{E}[Y - \widehat{Y}]^2}.$$

which immediately proves (5.26).

The Cholesky decomposition and the precision matrix

We now represent the precision matrix through its Cholesky decomposition. It should be mentioned that Mohsen Pourahmadi has done a lot of interesting research in this area and he recently wrote a review paper, which can be found here.

We define the sequence of linear equations

$$X_t = \sum_{j=1}^{t-1} \beta_{t,j} X_j + \varepsilon_t, \quad t = 2, \ldots, k, \tag{5.30}$$

where $\{\beta_{t,j}; 1 \leq j \leq t-1\}$ are the coefficeints of the best linear predictor of X_t given $X_1, \ldots, X_{t-1}$. Let $\sigma_t^2 = \mathrm{var}[\varepsilon_t] = \mathrm{E}[X_t - \sum_{j=1}^{\boldsymbol{t-1}} \beta_{t,j} X_j]^2$ and $\sigma_1^2 = \mathrm{var}[X_1]$. We standardize (5.30) and define

$$\sum_{j=1}^{t} \gamma_{t,j} X_j = \frac{1}{\sigma_t}\left(X_t - \sum_{j=1}^{\boldsymbol{t-1}} \beta_{t,j} X_j\right), \tag{5.31}$$

where we set $\gamma_{t,t} = 1/\sigma_t$ and for $1 \leq j < t-1$, $\gamma_{t,j} = -\beta_{t,j}/\sigma_i$. By construction it is clear that $\mathrm{var}(L\underline{X}) = I_k$, where

$$L = \begin{pmatrix} \gamma_{1,1} & 0 & 0 & \ldots & 0 & 0 \\ \gamma_{2,1} & \gamma_{2,2} & 0 & \ldots & 0 & 0 \\ \gamma_{3,1} & \gamma_{3,2} & \gamma_{3,3} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{k,1} & \gamma_{k,2} & \gamma_{k,3} & \ldots & \gamma_{k,k-1} & \gamma_{k,k} \end{pmatrix} \tag{5.32}$$

and $LL = \Sigma^{-1}$ (see Pourahmadi, equation (18)), where $\Sigma = \mathrm{var}(\boldsymbol{X}_k)$. Let $\Sigma = \mathrm{var}[\boldsymbol{X}_k]$, then

$$\Sigma^{ij} = \sum_{s=1}^{k} \gamma_{is}\gamma_{js} \qquad \text{(note many of the elements will be zero).}$$

Remark 5.5.2 (The Cholesky decomposition of a matrix) *All positive definite matrices admit a Cholesky decomposition. That is $H'H = Sigma$, where H is a lower triangular matrix. Similarly, $Sigma^{-1} = LL'$, where L is a lower triangular matrix and $L = H^{-1}$. Therefore we observe that if $\Sigma = \mathrm{var}(\underline{X})$ (where $\underline{X}$ is a p-dimension random vector), then*

$$\mathrm{var}\left(L\underline{X}\right) = L'\Sigma L = L'H'HL = I_p.$$

Therefore, the lower triangular matrix L "finds" a linear combination of the elements $\underline{X}$ such that the resulting random vector is uncorrelated.

We use apply these results to the analysis of the partial correlations of autoregressive processes and the inverse of its variance/covariance matrix.

A little bit more indepth: general vector spaces

First a brief definition of a vector space. $\mathcal{X}$ is called an vector space if for every $x, y \in \mathcal{X}$ and $a, b \in \mathbb{R}$ (this can be generalised to $\mathbb{C}$), then $ax + by \in \mathcal{X}$. An inner product space is a vector space which comes with an inner product, in other words for every element $x, y \in \mathcal{X}$ we can defined an innerproduct $\langle x, y\rangle$, where $\langle\cdot,\cdot\rangle$ satisfies all the conditions of an inner product. Thus for every element $x \in \mathcal{X}$ we can define its norm as $\|x\| = \langle x, x\rangle$. If the inner product space is complete (meaning the limit of every sequence in the space is also in the space) then the innerproduct space is a Hilbert space (see wiki).

Example 5.5.1 *(i) The Euclidean space $\mathbb{R}^n$ described above is a classical example of a Hilbert space. Here the innerproduct between two elements is simply the scalar product, $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^{n} x_i y_i$.*

(ii) The subset of the probability space $(\Omega, \mathcal{F}, P)$, where all the random variables defined on Ω have a finite second moment, ie. $\mathrm{E}(X^2) = \int_\Omega X(\omega)^2 dP(\omega) < \infty$. This space is denoted as $L^2(\Omega, \mathcal{F}, P)$. In this case, the inner product is $\langle X, Y\rangle = \mathrm{E}(XY)$.

(iii) The function space $L^2[\mathbb{R}, \mu]$, where $f \in L^2[\mathbb{R}, \mu]$ if f is mu-measureable and

$$\int_{\mathbb{R}} |f(x)|^2 d\mu(x) < \infty,$$

is a Hilbert space. For this space, the inner product is defined as

$$\langle f, g\rangle = \int_{\mathbb{R}} f(x)g(x)d\mu(x).$$

It is straightforward to generalize the above to complex random variables and functions defined on $\mathbb{C}$*. We simply need to remember to take conjugates when defining the innerproduct, ie.* $\langle X, Y\rangle = \mathrm{cov}(X, \overline{Y})$ *and* $\langle f, g\rangle = \int_{\mathbb{C}} f(z)\overline{g(z)}d\mu(z)$*.*

In this chapter our focus will be on certain spaces of random variables which have a finite variance.

Basis

The random variables $\{X_t, X_{t-1}, \ldots, X_1\}$ span the space $\mathcal{X}_t^1$ (denoted as $\overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots, X_1)$), if for every $Y \in \mathcal{X}_t^1$, there exists coefficients $\{a_j \in \mathbb{R}\}$ such that

$$Y = \sum_{j=1}^{t} a_j X_{t+1-j}. \qquad (5.33)$$

Moreover, $\overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots, X_1) = \mathcal{X}_t^1$ if for every $\{a_j \in \mathbb{R}\}$, $\sum_{j=1}^{t} a_j X_{t+1-j} \in \mathcal{X}_t^1$. We now define the basis of a vector space, which is closely related to the span. The random variables $\{X_t, \ldots, X_1\}$ form a basis of the space $\mathcal{X}_t^1$, if for every $Y \in \mathcal{X}_t^1$ we have a representation (5.33) and this representation is unique. More precisely, there does not exist another set of coefficients $\{\phi_j\}$ such that $Y = \sum_{j=1}^{t} \phi_j X_{t+1-j}$. For this reason, one can consider a basis as the minimal span, that is the smallest set of elements which can span a space.

Definition 5.5.1 (Projections) *The projection of the random variable* Y *onto the space spanned by* $\overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots, X_1)$ *(often denoted as* $P_{X_t, X_{t-1}, \ldots, X_1}(Y)$*) is defined as* $P_{X_t, X_{t-1}, \ldots, X_1}(Y) = \sum_{j=1}^{t} c_j X_{t+1-j}$*, where* $\{c_j\}$ *is chosen such that the difference* $Y - P_{(X_t, X_{t-1}, \ldots, X_1)}(Y_t)$ *is uncorrelated (orthogonal/perpendicular) to any element in* $\overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots, X_1)$*. In other words,* $P_{X_t, X_{t-1}, \ldots, X_1}(Y_t)$ *is the best linear predictor of* Y *given* $X_t, \ldots, X_1$*.*

Orthogonal basis

An orthogonal basis is a basis, where every element in the basis is orthogonal to every other element in the basis. It is straightforward to orthogonalize any given basis using the method of projections.

To simplify notation let $X_{t|t-1} = P_{X_{t-1},\ldots,X_1}(X_t)$. By definition, $X_t - X_{t|t-1}$ is orthogonal to the space $\overline{\text{sp}}(X_{t-1}, X_{t-1}, \ldots, X_1)$. In other words $X_t - X_{t|t-1}$ and X_s ($1 \leq s \leq t$) are orthogonal ($\text{cov}(X_s, (X_t - X_{t|t-1}))$, and by a similar argument $X_t - X_{t|t-1}$ and $X_s - X_{s|s-1}$ are orthogonal.

Thus by using projections we have created an orthogonal basis $X_1, (X_2 - X_{2|1}), \ldots, (X_t - X_{t|t-1})$ of the space $\overline{\text{sp}}(X_1, (X_2 - X_{2|1}), \ldots, (X_t - X_{t|t-1}))$. By construction it clear that $\overline{\text{sp}}(X_1, (X_2 - X_{2|1}), \ldots, (X_t - X_{t|t-1}))$ is a subspace of $\overline{\text{sp}}(X_t, \ldots, X_1)$. We now show that
$\overline{\text{sp}}(X_1, (X_2 - X_{2|1}), \ldots, (X_t - X_{t|t-1})) = \overline{\text{sp}}(X_t, \ldots, X_1)$.

To do this we define the sum of spaces. If U and V are two orthogonal vector spaces (which share the same innerproduct), then $y \in U \oplus V$, if there exists a $u \in U$ and $v \in V$ such that $y = u + v$. By the definition of $\mathcal{X}_t^1$, it is clear that $(X_t - X_{t|t-1}) \in \mathcal{X}_t^1$, but $(X_t - X_{t|t-1}) \notin \mathcal{X}_{t-1}^1$. Hence $\mathcal{X}_t^1 = \bar{sp}(X_t - X_{t|t-1}) \oplus \mathcal{X}_{t-1}^1$. Continuing this argument we see that $\mathcal{X}_t^1 = \bar{sp}(X_t - X_{t|t-1}) \oplus \bar{sp}(X_{t-1} - X_{t-1|t-2}) \oplus, \ldots, \oplus \bar{sp}(X_1)$. Hence $\bar{sp}(X_t, \ldots, X_1) = \bar{sp}(X_t - X_{t|t-1}, \ldots, X_2 - X_{2|1}, X_1)$. Therefore for every $P_{X_t,\ldots,X_1}(Y) = \sum_{j=1}^t a_j X_{t+1-j}$, there exists coefficients $\{b_j\}$ such that

$$P_{X_t,\ldots,X_1}(Y) = P_{X_t - X_{t|t-1},\ldots,X_2 - X_{2|1},X_1}(Y) = \sum_{j=1}^{t} P_{X_{t+1-j} - X_{t+1-j|t-j}}(Y) = \sum_{j=1}^{t-1} b_j (X_{t+1-j} - X_{t+1-j|t-j}) + b_t X_1,$$

where $b_j = \text{E}(Y(X_j - X_{j|j-1}))/\text{E}(X_j - X_{j|j-1}))^2$. A useful application of orthogonal basis is the ease of obtaining the coefficients b_j, which avoids the inversion of a matrix. This is the underlying idea behind the innovations algorithm proposed in Brockwell and Davis (1998), Chapter 5.

Spaces spanned by infinite number of elements (advanced)

The notions above can be generalised to spaces which have an infinite number of elements in their basis. Let now construct the space spanned by infinite number random variables $\{X_t, X_{t-1}, \ldots\}$. As with anything that involves ∞ we need to define precisely what we mean by an infinite basis. To do this we construct a sequence of subspaces, each defined with a finite number of elements in the basis. We increase the number of elements in the subspace and consider the limit of this space. Let $\mathcal{X}_t^{-n} = \overline{\text{sp}}(X_t, \ldots, X_{-n})$, clearly if $m > n$, then $\mathcal{X}_t^{-n} \subset \mathcal{X}_t^{-m}$. We define $X_t^{-\infty}$, as $X_t^{-\infty} = \cup_{n=1}^{\infty} \mathcal{X}_t^{-n}$, in other words if $Y \in \mathcal{X}_t^{-\infty}$, then there exists an n such that $Y \in \mathcal{X}_t^{-n}$. However, we also need to ensure that the limits of all the sequences lie in this infinite dimensional space, therefore we close the space by defining defining a new space which includes the old space and also includes all the limits. To make this precise suppose the sequence of random variables is such that $Y_s \in \mathcal{X}_t^{-s}$, and $\mathrm{E}(Y_{s_1} - Y_{s_2})^2 \to 0$ as $s_1, s_2 \to \infty$. Since the sequence $\{Y_s\}$ is a Cauchy sequence there exists a limit. More precisely, there exists a random variable Y, such that $\mathrm{E}(Y_s - Y)^2 \to 0$ as $s \to \infty$. Since the closure of the space, $\overline{\mathcal{X}}_t^{-n}$, contains the set $\mathcal{X}_t^{-n}$ and all the limits of the Cauchy sequences in this set, then $Y \in \overline{\mathcal{X}_t^{-\infty}}$. We let

$$\overline{\mathcal{X}_t^{-\infty}} = \overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots), \tag{5.34}$$

The orthogonal basis of $\overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots)$

An orthogonal basis of $\overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots)$ can be constructed using the same method used to orthogonalize $\overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots, X_1)$. The main difference is how to deal with the initial value, which in the case of $\overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots, X_1)$ is X_1. The analogous version of the initial value in infinite dimension space $\overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots)$ is $X_{-\infty}$, but this it not a well defined quantity (again we have to be careful with these pesky infinities).

Let $X_{t-1}(1)$ denote the best linear predictor of X_t given $X_{t-1}, X_{t-2}, \ldots$. As in Section 5.5 it is clear that $(X_t - X_{t-1}(1))$ and X_s for $s \leq t-1$ are uncorrelated and $\overline{X_t^{-\infty}} = \overline{\mathrm{sp}}(X_t - X_{t-1}(1)) \oplus \overline{X_{t-1}^{-\infty}}$, where $\overline{X_t^{-\infty}} = \overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots)$. Thus we can construct the orthogonal basis $(X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots$ and the corresponding space $\overline{\mathrm{sp}}((X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots)$. It is clear that $\overline{\mathrm{sp}}((X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots) \subset \overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots)$. However, unlike the finite dimensional case it is not clear that they are equal, roughly speaking this is because $\overline{\mathrm{sp}}((X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots)$ lacks the inital value $X_{-\infty}$. Of course the time $-\infty$ in the past is not really a well defined quantity. Instead, the way we overcome this issue is that we define the initial starting random variable as the intersection of the subspaces, more precisely let $\mathcal{X}_{-\infty} = \cap_{n=-\infty}^{\infty} \mathcal{X}_t^{-\infty}$. Furthermore, we note that since $X_n - X_{n-1}(1)$ and X_s (for any $s \leq n$) are orthogonal, then $\overline{\mathrm{sp}}((X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots)$ and $\mathcal{X}_{-\infty}$ are orthogonal spaces. Using $\mathcal{X}_{-\infty}$, we have $\oplus_{j=0}^{t} \overline{\mathrm{sp}}((X_{t-j} - X_{t-j-1}(1)) \oplus \mathcal{X}_{-\infty} = \overline{\mathrm{sp}}(X_t, X_{t-1}, \ldots)$.

Chapter 6

The autocovariance and partial covariance of a stationary time series

Objectives

- Be able to determine the rate of decay of an ARMA time series.
- Be able 'solve' the autocovariance structure of an AR process.
- Understand what partial correlation is and how this may be useful in determining the order of an AR model.

6.1 The autocovariance function

The autocovariance function (ACF) is defined as the sequence of covariances of a stationary process. Precisely, suppose $\{X_t\}$ is a stationary process with mean zero, then $\{c(r) : k \in \mathbb{Z}\}$ is the ACF of $\{X_t\}$ where $c(r) = \text{cov}(X_0, X_r)$. The autocorrelation function is the standardized version of the autocovariance and is defined as

$$\rho(r) = \frac{c(r)}{c(0)}.$$

Clearly different time series give rise to different features in the ACF. We will explore some of these features below.

Before investigating the structure of ARMA processes we state a general result connecting linear time series and the summability of the autocovariance function.

Lemma 6.1.1 *Suppose the stationary time series X_t satisfies the linear representation $\sum_{j=-\infty}^{\infty}\psi_j\varepsilon_{t-j}$. The covariance is $c(r) = \sum_{j=-\infty}^{\infty}\psi_j\psi_{j+r}$.*

(i) If $\sum_{j=\infty}^{\infty}|\psi_j| < \infty$, then $\sum_k |c(k)| < \infty$.

(ii) If $\sum_{j=\infty}^{\infty}|j\psi_j| < \infty$, then $\sum_k |k\cdot c(k)| < \infty$.

(iii) If $\sum_{j=\infty}^{\infty}|\psi_j|^2 < \infty$, then we cannot say anything about summability of the covariance.

PROOF. It is straightforward to show that

$$c(k) = \mathrm{var}[\varepsilon_t]\sum_j \psi_j\psi_{j-k}.$$

Using this result, it is easy to see that $\sum_k |c(k)| \leq \sum_k\sum_j |\psi_j|\cdot|\psi_{j-k}|$, thus $\sum_k |c(k)| < \infty$, which proves (i).

The proof of (ii) is similar. To prove (iii), we observe that $\sum_j |\psi_j|^2 < \infty$ is a weaker condition then $\sum_j |\psi_j| < \infty$ (for example the sequence $\psi_j = |j|^{-1}$ satisfies the former condition but not the latter). Thus based on the condition we cannot say anything about summability of the covariances. □

First we consider a general result on the covariance of a causal ARMA process (always to obtain the covariance we use the MA(∞) expansion - you will see why below).

6.1.1 The rate of decay of the autocovariance of an ARMA process

We evaluate the covariance of an ARMA process using its MA(∞) representation. Let us suppose that $\{X_t\}$ is a causal ARMA process, then it has the representation in (4.20) (where the roots of $\phi(z)$ have absolute value greater than $1+\delta$). Using (4.20) and the independence of $\{\varepsilon_t\}$ we have

$$\begin{aligned}\mathrm{cov}(X_t, X_\tau) &= \mathrm{cov}(\sum_{j_1=0}^{\infty}a_{j_1}\varepsilon_{t-j_1}, \sum_{j_2=0}^{\infty}a_{j_2}\varepsilon_{\tau-j_2}) \\ &= \sum_{j=0}^{\infty}a_{j_1}a_{j_2}\mathrm{cov}(\varepsilon_{t-j},\varepsilon_{\tau-j}) = \sum_{j=0}^{\infty}a_ja_{j+|t-\tau|}\mathrm{var}(\varepsilon_t) \qquad (6.1)\end{aligned}$$

(here use the MA(∞) expansion). Using (4.21) we have

$$|\mathrm{cov}(X_t, X_\tau)| \leq \mathrm{var}(\varepsilon_t) C_\rho^2 \sum_{j=0}^{\infty} \rho^j \rho^{j+|t-\tau|} \leq C_\rho^2 \rho^{|t-\tau|} \sum_{j=0}^{\infty} \rho^{2j} = C^2 \frac{\rho^{|t-\tau|}}{1-\rho^2}, \tag{6.2}$$

for any $1/(1+\delta) < \rho < 1$.

The above bound is useful, it tells us that the ACF of an ARMA process decays exponentially fast. In other words, there is very little memory in an ARMA process. However, it is not very enlightening about features within the process. In the following we obtain an explicit expression for the ACF of an autoregressive process. So far we have used the characteristic polynomial associated with an AR process to determine whether it was causal. Now we show that the roots of the characteristic polynomial also give information about the ACF and what a 'typical' realisation of a autoregressive process could look like.

6.1.2 The autocovariance of an autoregressive process and the Yule-Walker equations

Simple worked example Let us consider the two AR(1) processes considered in Section 4.3.2. We recall that the model

$$X_t = 0.5X_{t-1} + \varepsilon_t$$

has the stationary causal solution

$$X_t = \sum_{j=0}^{\infty} 0.5^j \varepsilon_{t-j}.$$

Assuming the innovations has variance one, the ACF of X_t is

$$c_X(0) = \frac{1}{1-0.5^2} \qquad c_X(k) = \frac{0.5^{|k|}}{1-0.5^2}$$

The corresponding autocorrelation is

$$\rho_X(k) = 0.5^{|k|}.$$

Let us consider the sister model

$$Y_t = 2Y_{t-1} + \varepsilon_t,$$

this has the noncausal stationary solution

$$Y_t = -\sum_{j=0}^{\infty}(0.5)^{j+1}\varepsilon_{t+j+1}.$$

Thus process has the ACF

$$c_Y(0) = \frac{0.5^2}{1-0.5^2} \qquad c_X(k) = \frac{0.5^{2+|k|}}{1-0.5^2}.$$

The corresponding autocorrelation is

$$\rho_X(k) = 0.5^{|k|}.$$

Comparing the two ACFs, both models have identical autocorrelation function.

Therefore, we observe an interesting feature, that the non-causal time series has the same correlation structure of its dual causal time series. For every non-causal time series there exists a causal time series with the same autocovariance function. The dual is easily constructed. If an autoregressive model has characteristic function $\phi(z) = 1 - \sum_{j=1}^{p}\phi_j z^j$ with roots $\lambda_1, \ldots, \lambda_p$. If all the roots lie inside the unit circle, then $\phi(z)$ corresponds to a non-causal time series. But by flipping the roots $\lambda_1^{-1}, \ldots, \lambda_p^{-1}$ all the roots now lie outside the unit circle. This means the characteristic polynomial corresponding to $\lambda_1^{-1}, \ldots, \lambda_p^{-1}$ leads to a causal AR(p) model (call this $\widetilde{\phi}(z)$). More over the characteristic polynomial of the AR(p) models associated with $\phi(z)$ and $\widetilde{\phi}(z)$ have the <u>same</u> autocorrelation function. They are duals. In summary, autocorrelation is 'blind' to non-causality.

<u>Another worked example</u> Consider the AR(2) model

$$X_t = 2r\cos(\theta)X_{t-1} - r^2 X_{t-2} + \varepsilon_t, \tag{6.3}$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance one. We assume $0 < r < 1$ (which imposes causality on the model). Note, that the non-casual case ($r > 1$) will have the same autocovariance as the causal case with r flipped to r^{-1}. The corresponding characteristic polynomial is $1-2r\cos(\theta)z+r^2z^2$, which has roots $r^{-1}\exp(\pm i\theta)$. By using (6.11), below, the ACF is

$$c(k) = r^{|k|}\left[C_1\exp(ik\theta)+\bar{C}_1\exp(-ik\theta)\right].$$

Setting $C_1 = a\exp(ib)$, then the above can be written as

$$c(k) = ar^{|k|}\left(\exp(i(b+k\theta))+\exp(-i(b+k\theta))\right) = 2ar^{|k|}\cos\left(k\theta+b\right), \qquad (6.4)$$

where the above follows from the fact that the sum of a complex number and its conjugate is two times the real part of the complex number.

Consider the AR(2) process

$$X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t, \qquad (6.5)$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance one. The corresponding characteristic polynomial is $1-1.5z+0,75z^2$, which has roots $\sqrt{4/3}\exp(i\pi/6)$. Using (6.4) the autocovariance function of $\{X_t\}$ is

$$c(k) = a(\sqrt{3/4})^{|k|}\cos\left(k\frac{\pi}{6}+b\right).$$

We see that the covariance decays at an exponential rate, but there is a periodicity within the decay. This means that observations separated by a lag $k=12$ are more closely correlated than other lags, this suggests a quasi-periodicity in the time series. The ACF of the process is given in Figure 6.1. Notice that it decays to zero (relatively fast) but it also undulates. A plot of a realisation of the time series is given in Figure 6.2, notice the quasi-periodicity of about $2\pi/12$. To measure the magnitude of the period we also give the corresponding periodogram in Figure 6.2. Observe a peak at the frequency about frequency $2\pi/12 \approx 0.52$. We now generalise the results in the above AR(1) and AR(2) examples. Let us consider the general AR(p) process

$$X_t = \sum_{j=1}^{p}\phi_j X_{t-j} + \varepsilon_t.$$

Suppose the roots of the corresponding characteristic polynomial are *distinct* and we split them

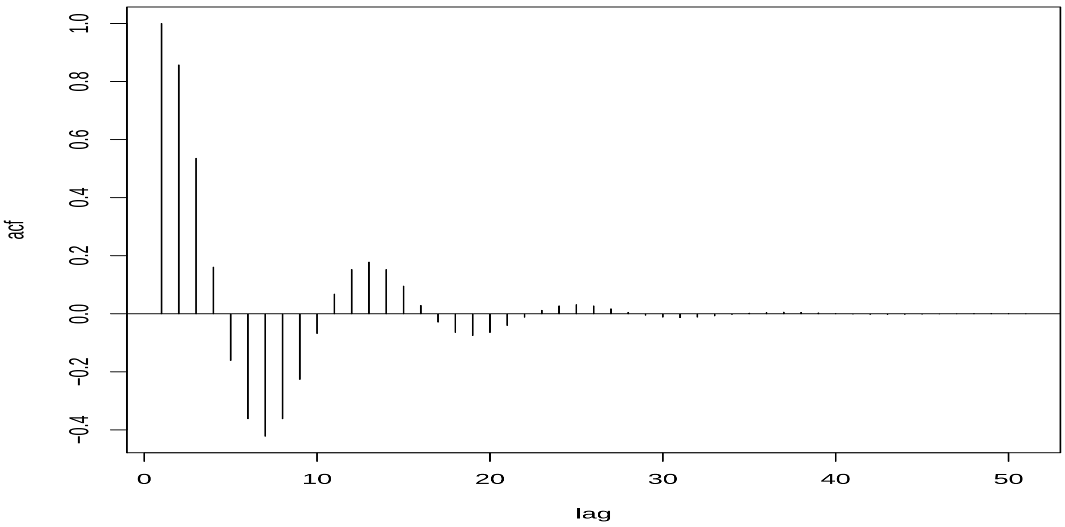


Figure 6.1: The ACF of the time series $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$

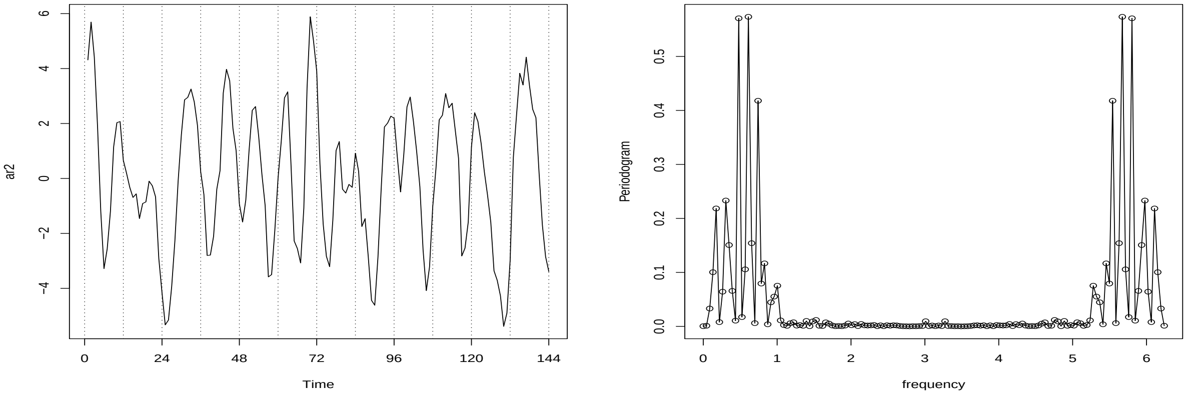


Figure 6.2: Left: A realisation from the time series $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$. Right: The corresponding periodogram.

into real and complex roots. Because the characteristic polynomial is comprised of real coefficients, the complex roots come in complex conjugate pairs. Hence let us suppose the real roots are $\{\lambda_j\}_{j=1}^{r}$ and the complex roots are $\{\lambda_j, \overline{\lambda}_j\}_{j=r+1}^{(p-r)/2}$. The covariance in (6.10) can be written as

$$c(k) = \sum_{j=1}^{r} C_j \lambda_j^{-k} + \sum_{j=r+1}^{(p-2)/2} a_j |\lambda_j|^{-k} \cos(k\theta_j + b_j)$$

where for $j > r$ we write $\lambda_j = |\lambda_j| \exp(i\theta_j)$ and a_j and b_j are real constants. Notice that as the example above the covariance decays exponentially with lag, but there is undulation. A typical realisation from such a process will be quasi-periodic with periods at $\theta_{r+1}, \ldots, \theta_{(p-r)/2}$, though the magnitude of each period will vary.

Exercise 6.1 *Recall the AR(2) models considered in Exercise 4.5. Now we want to derive their ACF functions.*

(i) *(a)* *Obtain the ACF corresponding to*

$$X_t = \frac{7}{3}X_{t-1} - \frac{2}{3}X_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2.

(b) *Obtain the ACF corresponding to*

$$X_t = \frac{4 \times \sqrt{3}}{5}X_{t-1} - \frac{4^2}{5^2}X_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2.

(c) *Obtain the ACF corresponding to*

$$X_t = X_{t-1} - 4X_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2.

(ii) *For all these models plot the true ACF in* R. *You will need to use the function* `ARMAacf`. *BEWARE of the ACF it gives for non-causal solutions. Find a method of plotting a causal solution in the non-causal case.*

Exercise 6.2 *In Exercise 4.6 you constructed a causal AR(2) process with period 17.*

Load Shumway and Stoffer's package `astsa` *into R (use the command* `install.packages("astsa")` *and then* `library("astsa")`*.*

Use the command `arma.spec` *to make a plot of the corresponding spectral density function. How does your periodogram compare with the 'true' spectral density function?*

Derivation of the ACF of general models (advanced)

Worked example Let us suppose that X_t satisfies the model $X_t = (a+b)X_{t-1} - abX_{t-2} + \varepsilon_t$. We have shown that if $|a| < 1$ and $|b| < 1$, then it has the solution

$$X_t = \frac{1}{b-a}\big(\sum_{j=0}^{\infty}\big(b^{j+1} - a^{j+1})\varepsilon_{t-j}\big).$$

By matching the innovations it can be shown that for $r > 0$

$$\text{cov}(X_t, X_{t+r}) \quad = \quad \sum_{j=0}^{\infty}(b^{j+1} - a^{j+1})(b^{j+1+r} - a^{j+1+r}). \tag{6.6}$$

Even by using the sum of a geometric series the above is still cumbersome. Below we derive the general solution, which can be easier to interprete.

General AR(p) models

Let us consider the zero mean AR(p) process $\{X_t\}$ where

$$X_t = \sum_{j=1}^{p}\phi_j X_{t-j} + \varepsilon_t. \tag{6.7}$$

From now onwards we will assume that $\{X_t\}$ is causal (the roots of $\phi(z)$ lie outside the unit circle). Evaluating the covariance of above with respect X_{t-k} ($k \leq 0$) gives the sequence of equations

$$\text{cov}(X_t X_{t-k}) \quad = \quad \sum_{j=1}^{p}\phi_j \text{cov}(X_{t-j}, X_{t-k}). \tag{6.8}$$

It is worth mentioning that if the process were not causal this equation would not hold, since ε_t and X_{t-k} are not uncorrelated. Let $c(r) = \text{cov}(X_0, X_r)$ and substituting into the above gives the sequence of difference equations

$$c(k) - \sum_{j=1}^{p}\phi_j c(k-j) = 0, \qquad k \geq 0. \tag{6.9}$$

The autocovariance function of $\{X_t\}$ is the solution of this difference equation. Solving (6.9) is very similar to solving homogenuous differential equations, which some of you may be familar with (do not worry if you are not).

Recall the characteristic polynomial of the AR process $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j = 0$, which has the roots $\lambda_1, \ldots, \lambda_p$. In Section 4.3.3 we used the roots of the characteristic equation to find the stationary solution of the AR process. In this section we use the roots characteristic to obtain the solution (6.9). We show below that if the roots are distinct (the roots are all different) the solution of (6.9) is

$$c(k) = \sum_{j=1}^{p} C_j \lambda_j^{-|k|}, \tag{6.10}$$

where the constants $\{C_j\}$ are chosen depending on the initial values $\{c(k) : 1 \leq k \leq p\}$. If λ_j is real, then C_j is real. If λ_j is complex, then it will have another root λ_{j+1}. Consequently, C_j and C_{j+1} will be complex conjugations of each other. This is to ensure that $\{c(k)\}_k$ is real.

Example $p = 2$ Suppose the roots of $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$ are complex (and this conjugates). Then

$$c(k) \quad = \quad C_1 \lambda_1^{-|k|} + C_2 \lambda_2^{-|k|} = C\lambda^{-|k|} + \overline{C\lambda}^{-|k|}. \tag{6.11}$$

Proof of (6.10) The simplest way to prove (6.10) is to use the plugin method (guess a solution and plug it in). Plugging $c(k) = \sum_{j=1}^{p} C_j \lambda_j^{-k}$ into (6.9) gives

$$\begin{aligned} c(k) - \sum_{j=1}^{p} \phi_j c(k-j) &= \sum_{j=1}^{p} C_j \left(\lambda_j^{-k} - \sum_{i=1}^{p} \phi_i \lambda_j^{-(k-i)} \right) \\ &= \sum_{j=1}^{p} C_j \lambda_j^{-k} \underbrace{\left(1 - \sum_{i=1}^{p} \phi_i \lambda_j^{i} \right)}_{\phi(\lambda_i)} = 0. \end{aligned}$$

which proves that it is a solution. □

Non-distinct roots In the case that the roots of $\phi(z)$ are not distinct, let the roots be $\lambda_1, \ldots, \lambda_s$ with multiplicity $m_1, \ldots, m_s$ ($\sum_{k=1}^{s} m_k = p$). In this case the solution is

$$c(k) = \sum_{j=1}^{s} \lambda_j^{-k} P_{m_j}(k),$$

where $P_{m_j}(k)$ is m_jth order polynomial and the coefficients $\{C_j\}$ are now 'hidden' in $P_{m_j}(k)$.

6.1.3 The autocovariance of a moving average process

Suppose that $\{X_t\}$ satisfies

$$X_t = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}.$$

The covariance is

$$\text{cov}(X_t, X_{t-k}) = \begin{cases} \sum_{i=0}^{p} \theta_i \theta_{i-k} & k = -q, \ldots, q \\ 0 & \text{otherwise} \end{cases}$$

where $\theta_0 = 1$ and $\theta_i = 0$ for $i < 0$ and $i \geq q$. Therefore we see that there is no correlation when the lag between X_t and X_{t-k} is greater than q.

6.1.4 The autocovariance of an ARMA process (advanced)

We see from the above that an MA(q) model is only really suitable when we believe that there is no correlaton between two random variables separated by more than a certain distance. Often autoregressive models are fitted. However in several applications we find that autoregressive models of a very high order are needed to fit the data. If a very 'long' autoregressive model is required a more suitable model may be the autoregressive moving average process. It has several of the properties of an autoregressive process, but can be more parsimonuous than a 'long' autoregressive process. In this section we consider the ACF of an ARMA process.

Let us suppose that the causal time series $\{X_t\}$ satisfies the equations

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}.$$

We now define a recursion for ACF, which is similar to the ACF recursion for AR processes. Let us suppose that the lag k is such that $k > q$, then it can be shown that the autocovariance function of the ARMA process satisfies

$$\text{cov}(X_t, X_{t-k}) - \sum_{i=1}^{p} \phi_i \text{cov}(X_{t-i}, X_{t-k}) = 0 \qquad k > q.$$

On the other hand, if $k \leq q$, then we have

$$\text{cov}(X_t, X_{t-k}) - \sum_{i=1}^{p} \phi_i \text{cov}(X_{t-i}, X_{t-k}) \;=\; \sum_{j=1}^{q} \theta_j \text{cov}(\varepsilon_{t-j}, X_{t-k}) = \sum_{j=k}^{q} \theta_j \text{cov}(\varepsilon_{t-j}, X_{t-k}).$$

We recall that X_t has the MA(∞) representation $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ (see (4.20)), therefore for $k \leq j \leq q$ we have $\text{cov}(\varepsilon_{t-j}, X_{t-k}) = a_{j-k}\text{var}(\varepsilon_t)$ (where $a(z) = \theta(z)\phi(z)^{-1}$). Altogether the above gives the difference equations

$$\begin{aligned} c(k) - \sum_{i=1}^{p} \phi_i c(k-i) &= \text{var}(\varepsilon_t) \sum_{j=k}^{q} \theta_j a_{j-k} \quad \text{for } 1 \leq k \leq q \\ c(k) - \sum_{i=1}^{p} \phi_i c(k-i) &= 0, \text{ for } k > q, \end{aligned}$$

where $c(k) = \text{cov}(X_0, X_k)$. Since the above is a is homogenuous difference equation, then it can be shown that the solution is

$$c(k) = \sum_{j=1}^{s} \lambda_j^{-k} P_{m_j}(k),$$

where $\lambda_1, \ldots, \lambda_s$ with multiplicity $m_1, \ldots, m_s$ ($\sum_k m_s = p$) are the roots of the characteristic polynomial $1 - \sum_{j=1}^{p} \phi_j z^j$. The coefficients in the polynomials P_{m_j} are determined by initial condition.

Further reading: Brockwell and Davis (1998), Chapter 3.3 and Shumway and Stoffer (2006), Chapter 3.4.

6.1.5 Estimating the ACF from data

Suppose we observe $\{Y_t\}_{t=1}^{n}$, to estimate the covariance we can estimate the covariance $c(k) = \text{cov}(Y_0, Y_k)$ from the the observations. One such estimator is

$$\widehat{c}_n(k) = \frac{1}{n} \sum_{t=1}^{n-|k|} (Y_t - \bar{Y}_n)(Y_{t+|k|} - \bar{Y}_n), \tag{6.12}$$

since $\mathrm{E}[(Y_t - \bar{Y}_n)(Y_{t+|k|} - \bar{Y}_n)] \approx c(k)$. Of course if the mean of Y_t is known to be zero ($Y_t = X_t$), then the simpler covariance estimator is

$$\widehat{c}_n(k) = \frac{1}{n}\sum_{t=1}^{n-|k|} X_t X_{t+|k|}.$$

The sample autocorrelation is the ratio

$$\widehat{\rho}_n(r) = \frac{\widehat{c}_n(r)}{\widehat{c}_n(0)}.$$

Thus for $r = 0$, we have $\widehat{\rho}_n(0) = 1$. Most statistical software will have functions that evaluate the sample autocorrelation function. In R, the standard function is `acf`. To illustrate the differences between the true ACF and estimated ACF (with sample size $n = 100$) we consider the model

$$X_t = 2 \cdot 0.9\cos(\pi/3)X_{t-1} - 0.9^2 X_{t-2} + \varepsilon_t.$$

We make a plot of the true ACF and estimated ACF in Figure **??**. As a contrast we consider the estimated and true ACF of the MA model

$$X_t = \varepsilon_t + 2 \cdot 0.9\cos(\pi/3)\varepsilon_{t-1} - 0.9^2\varepsilon_{t-2}. \tag{6.13}$$

This plot is given in Figure 6.4.

Observe that estimated autocorrelation plot contains a blue line. This blue line corresponds to $\pm 1.96/\sqrt{n}$ (where n is the sample size). These are the error bars, which are constructed under the assumption the data is actually iid. We show in Section 8.2 if $\{X_t\}$ are iid random variables then for all $h \geq 1$

$$\sqrt{n}\widehat{c}_n(h) \overset{\mathcal{D}}{\to} \mathcal{N}(0,1). \tag{6.14}$$

This gives rise to the critical values $\pm 1.96/\sqrt{n}$.

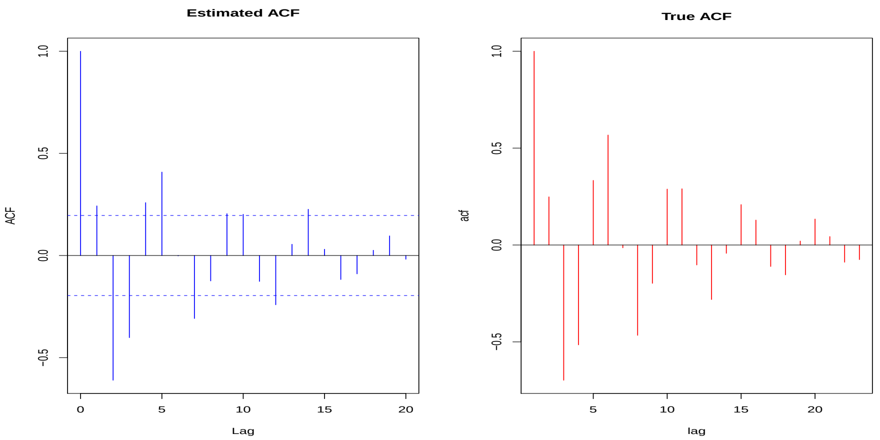


Figure 6.3: The AR(2) model. Left: Estimated ACF based on $n = 100$. Right: True ACF

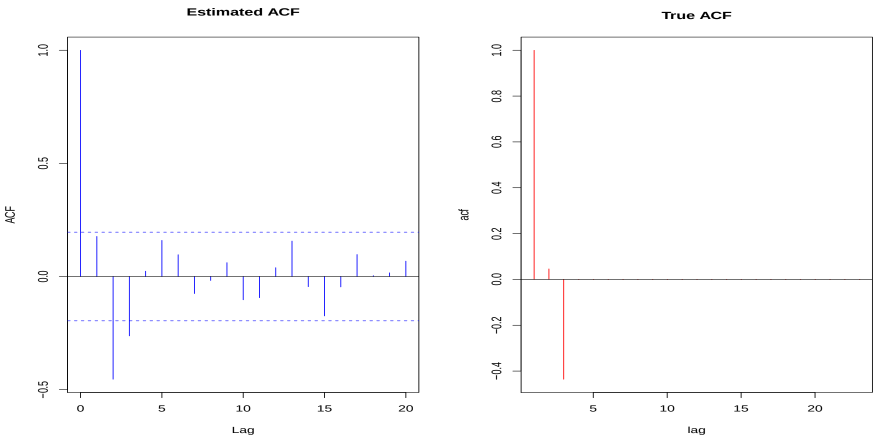


Figure 6.4: The MA(2) model. Left: Estimated ACF based on $n = 100$. Right: True ACF

6.2 Partial correlation in time series

6.2.1 A general definition

In Section 5.3 we introduced the notion of partial correlation for multivariate data. We now apply this notion to time series.

Definition 6.2.1 *Suppose that $\{X_t\}_t$ is a time series. The partial covariance/correlation between X_t and X_{t+k+1} is defined as the partial covariance/correlation between X_t and X_{t+k+1} after conditioning out the 'inbetween' time series $\underline{Y}' = (X_{t+1}, \ldots, X_{t+k})$. We denote this as $\rho_{t,t+k+1}(k)$,* *where*

$$\rho_k(t) = \frac{\text{cov}(X_t - P_{\underline{Y}}(X_t), X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1}))}{\sqrt{\text{var}(X_t - P_{\underline{Y}}(X_t))\text{var}(X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1}))}},$$

with

$$\begin{aligned}
& \text{cov}(X_t - P_{\underline{Y}}(X_t), X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1})) \\
= \; & \text{cov}(X_t, X_{t+k+1}) - \text{cov}(X_t, \underline{Y})'[\text{var}(\underline{Y})]^{-1}\text{cov}(X_{t+k+1}, \underline{Y}) \\
& \text{var}(X_t - P_{\underline{Y}}(X_t)) \\
= \; & \text{var}(X_t) - \text{cov}(X_t, \underline{Y})'[\text{var}(\underline{Y})]^{-1}\text{cov}(X_t, \underline{Y}) \\
& \text{var}(X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1})) \\
= \; & \text{var}(X_{t+k+1}) - \text{cov}(X_{t+k+1}, \underline{Y})'[\text{var}(\underline{Y})]^{-1}\text{cov}(X_{t+k+1}, \underline{Y}).
\end{aligned}$$

The above expression is horribly unwieldy. But many simplifications can be made once we impose the condition of second order stationarity.

6.2.2 Partial correlation of a stationary time series

If the time series is stationary, then the shift t becomes irrelevant (observe $\text{cov}(X_t, X_{t+k+1}) = c(k+1)$, $\text{cov}(X_t, X_t) = c(0)$ etc). We can center everything about $t = 0$, the only term that is relevant is the spacing k and define

$$\rho_{k+1|k+1} = \frac{\text{cov}(X_0 - P_{\underline{Y}}(X_0), X_{k+1} - P_{\underline{Y}}(X_{k+1}))}{\sqrt{\text{var}(X_0 - P_{\underline{Y}}(X_0))\text{var}(X_{k+1} - P_{\underline{Y}}(X_{k+1}))}},$$

where $\underline{Y}' = (X_1, X_2, \ldots, X_k)$,

$$\begin{aligned}
\text{cov}(X_t - P_{\underline{Y}}(X_t), X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1})) &= c(k+1) - \text{cov}(X_0, \underline{Y})'[\text{var}(\underline{Y})]^{-1}\text{cov}(X_{k+1}, \underline{Y}) \\
\text{var}(X_0 - P_{\underline{Y}}(X_0)) &= c(0) - \text{cov}(X_0, \underline{Y})'[\text{var}(\underline{Y})]^{-1}\text{cov}(X_0, \underline{Y}) \\
\text{var}(X_{k+1} - P_{\underline{Y}}(X_{k+1})) &= c(0) - \text{cov}(X_{k+1}, \underline{Y})'[\text{var}(\underline{Y})]^{-1}\text{cov}(X_{k+1}, \underline{Y}).
\end{aligned}$$

But there exists another interesting trick that will simplify the above. The value of the above expression is that given the autocovariance function, one can evaluate the above. However, this involves inverting matrices. Below we simplify the above expression even further, and in Section

7.5.1 we show how partial correlation can be evaluated without inverting any matrices. We first note that by stationarity

$$\begin{aligned} \text{cov}(X_0, \underline{Y}') &= (c(1), c(2) \ldots, c(k+1)) \\ \text{and cov}(X_{n+1}, \underline{Y}') &= (c(k+1), c(2) \ldots, c(1)). \end{aligned}$$

Thus the two vectors $\text{cov}(X_0, \underline{Y}')$ and $\text{cov}(X_{k+1}, \underline{Y}')$ are flips/swaps of each other. The flipping action can be done with a matrix transformation $\text{cov}(X_0, \underline{Y}) = E_k \text{cov}(X_{k+1}, \underline{Y})$ where

$$E_k = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & \vdots & 0 & 0 & 0 \end{pmatrix}.$$

We now describe some useful implications of this result.

<u>Time reversibility property of stationary time series</u> For stationary time series, predicting into the future and predicting into the past leads to the <u>same set</u> of prediction coefficients (they are just flipped round). More precisely, the projection of X_{k+1} onto the space spanned by $\underline{Y} = (X_1, X_2, \ldots, X_k)$, is the best linear predictor of X_{k+1} given $\boldsymbol{X}_k$. We will denote the projection of X_k onto the space spanned by $\underline{Y}' = (X_1, X_2, \ldots, X_k)$ as $P_{\underline{Y}}(X_{k+1})$. Thus

$$P_{\underline{Y}}(X_{k+1}) = \underline{Y}' \text{var}[\underline{Y}]^{-1} \text{cov}[X_{k+1}, \underline{Y}] = \underline{Y}' \Sigma_k^{-1} \underline{c}_k := \sum_{j=1}^{k} \phi_{k,j} X_{k+1-j},$$

where $\Sigma_k = \text{var}(\underline{Y})$ and $\underline{c}_k = \text{cov}(X_{k+1}, \underline{Y})$. But by flipping/swapping the coefficients, the same construction can be used to predict into the past X_0:

$$P_{\underline{Y}}(X_0) = \sum_{j=1}^{k} \phi_{k,j} X_j = \sum_{j=1}^{k} \phi_{k,k+1-j} X_{k+1-j}. \tag{6.15}$$

<u>Proof of equation (6.15)</u>

$$P_{\underline{Y}}(X_0) = \underline{Y}' (\text{var}[\underline{Y}]^{-1} \text{cov}[X_0, \underline{Y}]).$$

However, second order stationarity implies that $\text{cov}[X_0, \underline{Y}]) = E_k\text{cov}[X_{k+1}, \underline{Y}]) = E_k\underline{c}_k$ Thus

$$\begin{aligned} P_{\underline{Y}}(X_0) &= (\Sigma_k^{-1}E_k\text{cov}[X_{k+1}, \underline{Y}]) \\ &= \underline{Y}'\Sigma_k^{-1}E_k\underline{c}_k = \underline{Y}'E_k\Sigma_k^{-1}\underline{c}_k := \sum_{j=1}^{k}\phi_{k,k+1-j}X_{k+1-j}. \end{aligned}$$

Thus proving (6.15). □

With a little thought, we realize the partial correlation between X_t and X_{t+k} (where $k > 0$) is the correlation $X_0 - P_{\underline{Y}}(X_0) = X_0 - \sum_{j=1}^{k}\phi_{k,j}X_j$ and $X_{k+1} - P_{\underline{Y}}(X_{k+1}) = X_{k+1} - \sum_{j=1}^{k}\phi_{k,j}X_{k+1-j}$, some algebra gives

$$\begin{aligned} \text{cov}(X_t - P_{\underline{Y}}(X_t), X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1})) &= c(0) - \underline{c}_k'E_k\Sigma_k^{-1}\underline{c}_k \\ \text{var}(X_0 - P_{\underline{Y}}(X_0)) &= \text{var}(X_{k+1} - P_{\underline{Y}}(X_{k+1})) = \text{var}(X_0) - \underline{c}_k'\Sigma_k^{-1}\underline{c}_k. \end{aligned}$$

The last line of the above is important. It states that the variance of the prediction error in the past $X_0 - P_{\underline{Y}}(X_0)$ has the same as the variance of the prediction error into the future $X_{k+1} - P_{\underline{Y}}(X_{k+1})$. This is because the process is stationary.

Thus the partial correlation is

$$\rho_{k+1|k_1} = \frac{c(k+1) - \underline{c}_k'E_k\Sigma_k^{-1}\underline{c}_k}{c(0) - \underline{c}_k'\Sigma_k^{-1}\underline{c}_k}. \tag{6.16}$$

In the section below we show that $\rho_{k+1|k+1}$ can be expressed in terms of the best fitting AR$(k+1)$ parameters (which we will first have to define).

6.2.3 Best fitting AR(p) model

So far we have discussed time series which is generated with an AR(2). But we have not discussed fitting an AR(p) model to any stationary time series (not necessarily where the true underlying data generating mechanism is an AR(p)), which is possibly more important. We will show that the partial correlation is related to these fitted parameters. We state precisely what we mean below.

Suppose that the stationary time series is $g$enuinely generated with the causal AR(p) model

$$X_t = \sum_{j=1}^{p}\phi_jX_{t-j} + \varepsilon_t \tag{6.17}$$

where $\{\varepsilon_t\}$ are iid random variables. Then the projection of X_t onto $\underline{Y} = (X_{t-p}, \ldots, X_{t-1})$ is

$$P_{\underline{Y}}(X_t) = \sum_{j=1}^{p} \phi_j X_{t-j}.$$

Since $\underline{Y}$ does not contain any (linear information) about the innovations $\{\varepsilon_t\}_t$. This means that $\{X_{t-j}\}_{j=1}^{p}$ are independent of ε_t. However, because (6.17) is the true model which generates the data, ε_t is independent of all $\{X_{t-j}\}$ for $j \geq 1$. But this is by virtue of the model and not the projection. The project can only ensure that $X_t - P_{\underline{Y}}(X_t)$ and $\underline{Y}$ are uncorrelated.

The best fitting AR(p) Now let us suppose that $\{X_t\}$ is a general second order stationary time series with autocovariance $\{c(r)\}_r$. We consider the projection of X_t onto $\underline{Y} = (X_{t-p}, \ldots, X_{t-1})$ (technically onto $\text{sp}(X_1, \ldots, X_n)$) this is

$$P_{\underline{Y}}(X_t) = \sum_{j=1}^{p} \phi_{p,j} X_{t-j}.$$

By construction $X_t - P_{\underline{Y}}(X_t)$ and $\underline{Y}$ are uncorrelated but $X_t - P_{\underline{Y}}(X_t)$ is not necessarily uncorrelated with $\{X_{t-j}\}$ for $j \geq (p+1)$. We call $\{\phi_{p,j}\}$ the best fitting AR(p) coefficients, because if the true model were an AR(p) model $\phi_{p,j} = \phi_j$. The best fitting AR(p) model is very important in applications. It is often used to forecast the time series into the future. Note we have already alluded to $\sum_{j=1}^{p} \phi_{p,j} X_{t-j}$ in the previous section. And we summarize these results again. Since $\sum_{j=1}^{p} \phi_{p,j} X_{t-j}$ is a projection onto $\underline{Y}$, the coefficients $\{\phi_{p,j}\}_{j=1}^{p}$ are

$$\underline{\phi}_p = [\text{var}(Y)]^{-1} \text{cov}(X_t, \underline{Y}) = \Sigma_p^{-1} \underline{c}_p,$$

where $[\Sigma_p]_{t,\tau} = c(t-\tau)$ and $\underline{c}_p' = (c(1), c(2), \ldots, c(p))$ (observe stationarity means these are invariant to shift).

6.2.4 Best fitting AR(p) parameters and partial correlation

We now state the main result which connects the best fitting AR(p) parameters with partial correlation. The partial correlation at lag $(p+1)$ is the last best fitting AR(p) coefficient $\phi_{p+1,p+1}$. More precisely

$$\rho_{p+1|p+1} = \phi_{p+1,p+1}. \tag{6.18}$$

It is this identity that is used to calculate (from the true ACF) and estimate (from the estimated ACF) partial correlation (and not the identity in (6.16), which is more cumbersome).

Proof of identity (6.18) To prove this result. We return to the classical multivariate case (in Section 5.3). In particular the identity (5.12) which relates the regression coefficients to the partial correlation:

$$\rho_{p+1|p+1} = \phi_{p+1|p+1}\sqrt{\frac{\mathrm{var}(\varepsilon_{0|X_1,\ldots,X_{p+1}})}{\mathrm{var}(\varepsilon_{p+1|X_0,\ldots,X_p})}}$$

where

$$\varepsilon_{0|X_1,\ldots,X_{p+1}} = X_0 - P_{X_1,\ldots,X_{p+1}}(X_0) \text{ and } \varepsilon_{p+1|X_0,\ldots,X_p} = X_{p+1} - P_{X_0,\ldots,X_p}(X_{p+1}).$$

Now the important observation. We recall from the previous section that the variance of the prediction error in the past, $X_0 - P_{X_1,\ldots,X_{p+1}}(X_0)$ is the same as the variance of the prediction error into the future, $X_{p+1} - P_{X_0,\ldots,X_p}(X_{p+1})$. Therefore $\mathrm{var}(\varepsilon_{0|X_1,\ldots,X_{p+1}}) = \mathrm{var}(\varepsilon_{p+1|X_0,\ldots,X_p})$ and

$$\rho_{p+1|p+1} = \phi_{p+1|p+1}.$$

This proves equation (6.18). □

Important observation Relating the AR(p) model to the partial correlations

Suppose the true data generating process is an AR(p_0), and we fit an AR(p) model to the data. If $p < p_0$, then

$$P_{X_{t-p},\ldots,X_{t-1}}(X_t) = \sum_{j=1}^{p}\phi_{p,j}X_{t-j}.$$

and $\rho_{p|p} = \phi_{p,p}$. If $p = p_0$, then

$$P_{X_{t-p_0},\ldots,X_{t-1}}(X_t) = \sum_{j=1}^{p_0}\phi_j X_{t-j}$$

and $\phi_{p_0,p_0} = \rho_{p_0} = \phi_{p_0}$. For any $p > p_0$, we have

$$P_{X_{t-p},\ldots,X_{t-1}}(X_t) = \sum_{j=1}^{p_0}\phi_j X_{t-j}.$$

Thus the coefficient is $\rho_{p|p} = \phi_{p,p} = 0$.

Thus for AR(p) models, the partial correlation of order greater than p will be zero. We visualize this property in the plots in the following section.

6.2.5 The partial autocorrelation plot

Of course given the time series $\{X_t\}_{t=1}^{n}$ the true partial correlation is unknown. Instead it is estimated from the data. This is done by sequentially fitting an AR(p) model of increasing order to the time series and extracting the parameter estimator $\widehat{\phi}_{p+1,p+1} = \widehat{\rho}_{p|p}$ and plotting $\widehat{\rho}_{p|p}$ against p. To illustrate the differences between the true ACF and estimated ACF (with sample size $n = 100$) we consider the model

$$X_t = 2 \cdot 0.9 \cos(\pi/3) X_{t-1} - 0.9^2 X_{t-2} + \varepsilon_t.$$

The empirical partial estimated partial autocorrelation plot ($n = 100$) and true correlation is given in Figures 6.5. As a contrast we consider the estimated ($n = 100$) and true ACF of the MA model

$$X_t = \varepsilon_t + 2 \cdot 0.9 \cos(\pi/3) \varepsilon_{t-1} - 0.9^2 \varepsilon_{t-2}.$$

The plot is given in Figure 6.6.

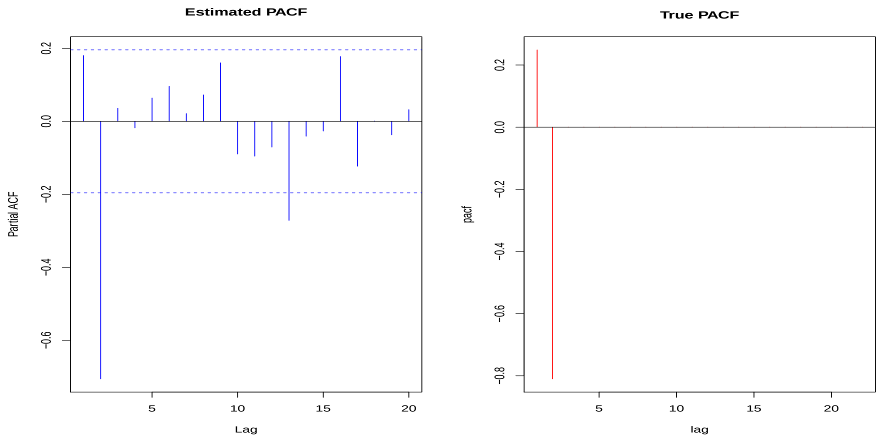


Figure 6.5: The AR(2): Left Estimated PACF ($n = 100$). Right: True PACF plot. $n = 100$

Observe that the partial correlation plot contains a blue line. This blue line corresponds to $\pm 1.96/\sqrt{n}$ (where n is the sample size).

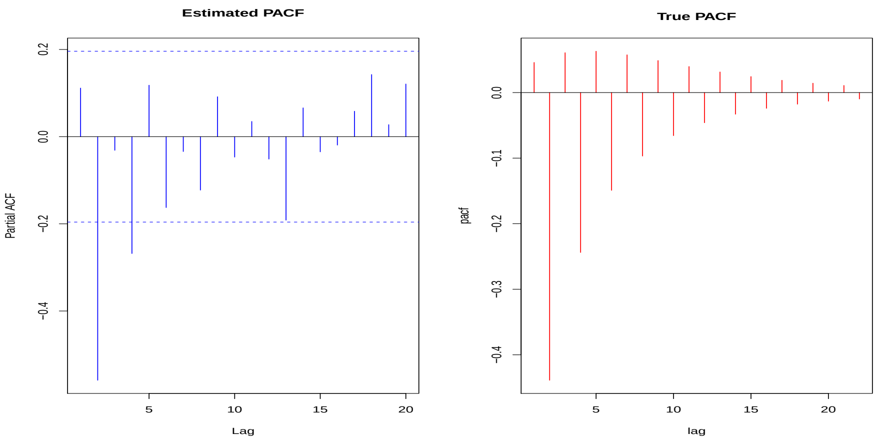


Figure 6.6: The MA(2): Left Estimated PACF ($n = 100$). Right: True PACF plot. $n = 100$

This blue line can be used as an aid in selecting the Autoregressive order (under certain conditions on the time series). We show in the next lecture that if $\{X_t\}$ is a *linear* time series with an AR(p) representation, then for $h > p$

$$\sqrt{n}\widehat{\rho}_{h|h} \overset{\mathcal{D}}{\to} \mathcal{N}(0,1), \tag{6.19}$$

which gives the critical values $\pm 1.96/\sqrt{n}$. But do not get too excited. We show that this result does not necessarily hold for non-linear time series. More precisely, the distribution will not be asymptotically pivotal.

6.2.6 Using the ACF and PACF for model identification

Figures 6.3, 6.4, 6.5 and 6.6 are very useful in identifying the model. We describe what we should observe below.

Using the ACF for model identification

If the true autocovariances after a certain lag are zero q, it may be appropriate to fit an MA(q) model to the time series. The $[-1.96n^{-1/2}, 1.96n^{-1/2}]$ error bars for an ACF plot *cannot* be reliably used to determine the order of an MA(q) model.

On the other hand, the autocovariances of any AR(p) process will only decay to zero as the lag increases (it will not be zero after a certain number of lags).

Using the PACF for model identification

If the true partial autocovariances after a certain lag are zero p, it may be appropriate to fit an AR(p) model to the time series.

Of course, in practice we only have the estimated partial autocorrelation at hand and not the true one. This is why we require the error bars. In Section 8.4 we show how these error bars are derived. The surprisingly result is that the error bars of a PACF can be used to determine the order of an AR(p) process. If the order of the autoregressive process is p, then for lag $r > p$, the partial correlation is such that $\widehat{\phi}_{rr} = N(0, n^{-1/2})$ (thus giving rise to the $[-1.96n^{-1/2}, 1.96n^{-1/2}]$ error bars). But It should be noted that there will be correlation between the sample partial correlations.

Exercise 6.3 (The partial correlation of an invertible MA(1)) *Let $\phi_{t,t}$ denote the partial correlation between X_{t+1} and X_1. It is well known (this is the Levinson-Durbin algorithm, which we cover in Chapter 7) that $\phi_{t,t}$ can be deduced recursively from the autocovariance funciton using the algorithm:*

Step 1 $\phi_{1,1} = c(1)/c(0)$ *and* $r(2) = \mathrm{E}[X_2 - X_{2|1}]^2 = \mathrm{E}[X_2 - \phi_{1,1}X_1]^2 = c(0) - \phi_{1,1}c(1)$.

Step 2 *For* $j = t$

$$\begin{aligned}
\phi_{t,t} &= \frac{c(t) - \sum_{j=1}^{t-1}\phi_{t-1,j}c(t-j)}{r(t)} \\
\phi_{t,j} &= \phi_{t-1,j} - \phi_{t,t}\phi_{t-1,t-j} \qquad 1 \leq j \leq t-1, \\
\text{and } r(t+1) &= r(t)(1-\phi_{t,t}^2).
\end{aligned}$$

(i) Using this algorithm and induction to show that the PACF of the MA(1) process $X_t = \varepsilon_t + \theta\varepsilon_{t-1}$, where $|\theta| < 1$ (so it is invertible) is

$$\phi_{t,t} = \frac{(-1)^{t+1}(\theta)^t(1-\theta^2)}{1-\theta^{2(t+1)}}.$$

Exercise 6.4 (Comparing the ACF and PACF of an AR process) *Compare the below plots:*

(i) Compare the ACF and PACF of the AR(2) model $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$ using `ARIMAacf(ar=c(1.5,-0.75),ma=0,30)` *and* `ARIMAacf(ar=c(1.5,-0.75),ma=0,pacf=T,30)`.

(ii) Compare the ACF and PACF of the MA(1) model $X_t = \varepsilon_t - 0.5\varepsilon_t$ using `ARIMAacf(ar=0,ma=c(-1.5),30)` *and* `ARIMAacf(ar=0,ma=c(-1.5),pacf=T,30)`.

(ii) Compare the ACF and PACF of the ARMA$(2,1)$ model $X_t - 1.5X_{t-1} + 0.75X_{t-2} = \varepsilon_t - 0.5\varepsilon_t$ using `ARIMAacf(ar=c(1.5,-0.75),ma=c(-1.5),30)` *and* `ARIMAacf(ar=c(1.5,0.75),ma=c(-1.5),pacf=T,30)`.

Exercise 6.5 *Compare the ACF and PACF plots of the monthly temperature data from 1996-2014. Would you fit an AR, MA or ARMA model to this data?*

`Rcode`

The sample partial autocorrelation of a time series can be obtained using the command `pacf`. However, remember just because the sample PACF is not zero, does not mean the true PACF is non-zero.

6.3 The variance and precision matrix of a stationary time series

Let us suppose that $\{X_t\}$ is a stationary time series. In this section we consider the variance/covariance matrix $\mathrm{var}(\underline{X}_n) = \Sigma_k$, where $\boldsymbol{X}_n = (X_1, \ldots, X_n)'$. We will consider two cases (i) when X_t follows an MA(p) models and (ii) when X_t follows an AR(p) model. The variance and inverse of the variance matrices for both cases yield quite interesting results. We will use classical results from multivariate analysis, stated in Chapter 5.

We recall that the variance/covariance matrix of a stationary time series has a (symmetric) Toeplitz structure (see wiki for a definition). Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)'$, then

$$\Sigma_n = \mathrm{var}(\boldsymbol{X}_n) = \begin{pmatrix} c(0) & c(1) & 0 & \ldots & c(n-2) & c(n-1) \\ c(1) & c(0) & c(1) & \ldots & c(n-3) & c(n-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ c(n-1) & c(n-2) & \vdots & \ldots & c(1) & c(0) \end{pmatrix}.$$

6.3.1 Variance matrix for AR(p) and MA(p) models

(i) If $\{X_t\}$ satisfies an MA(p) model and $n > p$, then Σ_n will be bandlimited, where p off-diagonals above and below the diagonal will be non-zero and the rest of the off-diagonal will be zero.

(ii) If $\{X_t\}$ satisfies an AR(p) model, then Σ_n will not be bandlimited.

Precision matrix for AR(p) models

We now consider the inverse of Σ_n. Warning: note that the inverse of a Toeplitz is not necessarily Toeplitz. Suppose that the time series $\{X_t\}_t$ has a causal AR(p) representation:

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid random variables with (for simplicity) variance $\sigma^2 = 1$. Let $\underline{X}_n = (X_1, \ldots, X_n)$ and suppose $n > p$.

Important result The inverse variance matrix Σ_n^{-1} is banded, with n non-zero bands off the diagonal.

Proof of claim We use the results in Chapter 5. Suppose that we have an AR(p) process and we consider the precision matrix of $\underline{X}_n = (X_1, \ldots, X_n)$, where $n > p$. To show this we use the Cholesky decomposition given in (5.30). This is where

$$\Sigma_n^{-1} = L_n L_n'$$

where L_n is the lower triangular matrix:

$$L_k = \begin{pmatrix} \phi_{1,0} & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \phi_{2,1} & \phi_{2,0} & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\phi_{p,p} & -\phi_{p,p-1} & \ldots & -\phi_{p,1} & \phi_{p,0} & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\phi_{n,n} & -\phi_{n,n-1} & \ldots & \ldots & & \ldots & -\phi_{n,4} & -\phi_{n,3} & -\phi_{n,2} & \phi_{n,1} & \phi_{n,0} \end{pmatrix} \tag{6.20}$$

where $\{\phi_{\ell,j}\}_{j=1}^{\ell}$ are the coefficients of the best linear predictor of X_ℓ given $\{X_{\ell-j}\}_{j=1}^{\ell-1}$ (after standardising by the residual variance). Since X_t is an autoregressive process of order p, if $t > p$, then

$$\phi_{t,j} = \begin{cases} \phi_j & 1 \leq j \leq p \\ 0 & j > p \end{cases}$$

This gives the lower triangular p-bandlimited matrix

$$L_n = \begin{pmatrix} \gamma_{1,0} & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ -\gamma_{2,1} & \gamma_{2,0} & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\phi_p & -\phi_{p-1} & \ldots & -\phi_1 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & -\phi_p & -\phi_{p-1} & \ldots & -\phi_1 & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & -\phi_p & \ldots & -\phi_2 & -\phi_1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{pmatrix}. \qquad (6.21)$$

Observe the above lower triangular matrix is zero after the pth off-diagonal.

Since $\Sigma_n^{-1} = L_n L_n'$ and L_n is a p-bandlimited matrix, $\Sigma_n^{-1} = L_n L_n'$ is a bandlimited matrix with the p off-diagonals either side of the diagonal non-zero. Let Σ^{ij} denote the (i,j)th element of Σ_k^{-1}. Then we observe that $\Sigma^{(i,j)} = 0$ if $|i-j| > p$. Moreover, if $0 < |i-j| \leq p$ and either i or j is greater than p. Further, from Section 5.4 we observe that the coefficients $\Sigma^{(i,j)}$ are the regression coefficients of X_i (after accounting for MSE).

Exercise 6.6 *Suppose that the time series $\{X_t\}$ has the causal $AR(2)$ representation*

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t.$$

Let $\underline{X}_n' = (X_1, \ldots, X_n)$ and $\Sigma_n = \mathrm{var}(\underline{X}_n)$. Suppose $L_n L_n' = \Sigma_n^{-1}$, where L_n is a lower triangular matrix.

(i) What does L_n looks like?

(ii) Using L_n evaluate the projection of X_t onto the space spanned by $\{X_{t-j}\}_{j\neq 0}$.

Remark 6.3.1 *Suppose that X_t is an autoregressive process $X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t$ where $\mathrm{var}[\varepsilon_t] = \sigma^2$ and $\{\varepsilon_t\}$ are uncorrelated random variables with zero mean. Let $\Sigma_m = \mathrm{var}[\boldsymbol{X}_m]$ where $\boldsymbol{X}_m = (X_1, \ldots, X_m)$. If $m > p$ then*

$$\left[\Sigma_m^{-1}\right]_{mm} = \Sigma^{mm} = \sigma^{-2}$$

and $\det(\Sigma_m) = \det(\Sigma_p)\sigma^{2(m-p)}$.

Exercise 6.7 *Prove Remark 6.3.1.*

6.4 The ACF of non-causal time series (advanced)

Here we demonstrate that it is not possible to identify whether a process is noninvertible/noncausal from its covariance structure. The simplest way to show result this uses the spectral density function, which will now define and then return to and study in depth in Chapter 10.

Definition 6.4.1 (The spectral density) *Given the covariances $c(k)$ (with $\sum_k |c(k)|^2 < \infty$) the spectral density function is defined as*

$$f(\omega) = \sum_k c(k) \exp(ik\omega).$$

The covariances can be obtained from the spectral density by using the inverse fourier transform

$$c(k) = \frac{1}{2\pi} \int_0^{2\pi} f(\omega) \exp(-ik\omega).$$

Hence the covariance yields the spectral density and visa-versa.

For reference below, we point out that the spectral density function uniquely identifies the autocovariance function.

Let us suppose that $\{X_t\}$ satisfies the AR(p) representation

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$$

where $\mathrm{var}(\varepsilon_t) = 1$ and the roots of $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j$ can lie inside and outside the unit circle, but not on the unit circle (thus it has a stationary solution). We will show in Chapter 10 that the spectral density of this AR process is

$$f(\omega) = \frac{1}{|1-\sum_{j=1}^{p}\phi_j \exp(ij\omega)|^2}. \tag{6.22}$$

- Factorizing $f(\omega)$.

  Let us supose the roots of the characteristic polynomial $\phi(z) = 1+\sum_{j=1}^{q}\phi_j z^j$ are $\{\lambda_j\}_{j=1}^{p}$, thus we can factorize $\phi(x)$ $1+\sum_{j=1}^{p}\phi_j z^j = \prod_{j=1}^{p}(1-\lambda_j z)$. Using this factorization we have (6.22) can be written as

$$f(\omega) \;=\; \frac{1}{\prod_{j=1}^{p}|1-\lambda_j \exp(i\omega)|^2}. \tag{6.23}$$

  As we have not assumed $\{X_t\}$ is causal, the roots of $\phi(z)$ can lie both inside and outside the unit circle. We separate the roots, into those outside the unit circle $\{\lambda_{O,j_1}; j_1 = 1,\ldots,p_1\}$ and inside the unit circle $\{\lambda_{I,j_2}; j_2 = 1,\ldots,p_2\}$ $(p_1+p_2 = p)$. Thus

$$\begin{aligned}\phi(z) &= [\prod_{j_1=1}^{p_1}(1-\lambda_{O,j_1}z)][\prod_{j_2=1}^{p_2}(1-\lambda_{I,j_2}z)] \\ &= (-1)^{p_2}\lambda_{I,j_2}z^{-p_2}[\prod_{j_1=1}^{p_1}(1-\lambda_{O,j_1}z)][\prod_{j_2=1}^{p_2}(1-\lambda_{I,j_2}^{-1}z)]. \end{aligned}\tag{6.24}$$

  Thus we can rewrite the spectral density in (6.25)

$$f(\omega) \;=\; \frac{1}{\prod_{j_2=1}^{p_2}|\lambda_{I,j_2}|^2}\frac{1}{\prod_{j_1=1}^{p_1}|1-\lambda_{O,j}\exp(i\omega)|^2\prod_{j_2=1}^{p_2}|1-\lambda_{I,j_2}^{-1}\exp(i\omega)|^2}. \tag{6.25}$$

  Let

$$f_O(\omega) \;=\; \frac{1}{\prod_{j_1=1}^{p_1}|1-\lambda_{O,j}\exp(i\omega)|^2\prod_{j_2=1}^{p_2}|1-\lambda_{I,j_2}^{-1}\exp(i\omega)|^2}.$$

  Then $f(\omega) = \prod_{j_2=1}^{p_2}|\lambda_{I,j_2}|^{-2}f_O(\omega)$.

- A parallel causal AR(p) process with the same covariance structure always exists.

  We now define a process which has the same autocovariance function as $\{X_t\}$ but is causal.

Using (6.24) we define the polynomial

$$\widetilde{\phi}(z) = [\prod_{j_1=1}^{p_1}(1-\lambda_{O,j_1}z)][\prod_{j_2=1}^{p_2}(1-\lambda_{I,j_2}^{-1}z)]. \tag{6.26}$$

By construction, the roots of this polynomial lie outside the unit circle. We then define the AR(p) process

$$\widetilde{\phi}(B)\widetilde{X}_t = \varepsilon_t, \tag{6.27}$$

from Lemma 4.3.1 we know that $\{\widetilde{X}_t\}$ has a stationary, almost sure unique solution. Moreover, because the roots lie outside the unit circle the solution is causal.

By using (6.22) the spectral density of $\{\widetilde{X}_t\}$ is $\widetilde{f}(\omega)$. We know that the spectral density function uniquely gives the autocovariance function. Comparing the spectral density of $\{\widetilde{X}_t\}$ with the spectral density of $\{X_t\}$ we see that they both are the same up to a multiplicative constant. Thus they both have the same autocovariance structure up to a multiplicative constant (which can be made the same, if in the definition (6.27) the innovation process has variance $\prod_{j_2=1}^{p_2}|\lambda_{I,j_2}|^{-2}$).

Therefore, for every non-causal process, there exists a causal process with the same autocovariance function.

By using the same arguments above, we can generalize to result to ARMA processes.

Definition 6.4.2 *An ARMA process is said to have minimum phase when the roots of $\phi(z)$ and $\theta(z)$ both lie outside of the unit circle.*

Remark 6.4.1 *For Gaussian random processes it is impossible to discriminate between a causal and non-causal time series, this is because the mean and autocovariance function uniquely identify the process.*

However, if the innovations are non-Gaussian, even though the autocovariance function is 'blind' to non-causal processes, by looking for other features in the time series we are able to discriminate between a causal and non-causal process.

6.4.1 The Yule-Walker equations of a non-causal process

Once again let us consider the zero mean AR(p) model

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t,$$

and $\mathrm{var}(\varepsilon_t) < \infty$. Suppose the roots of the corresponding characteristic polynomial lie outside the unit circle, then $\{X_t\}$ is strictly stationary where the solution of X_t is only in terms of past and present values of $\{\varepsilon_t\}$. Moreover, it is second order stationary with covariance $\{c(k)\}$. We recall from Section 6.1.2, equation (6.8) that we derived the Yule-Walker equations for causal AR(p) processes, where

$$\mathrm{E}(X_t X_{t-k}) \quad = \quad \sum_{j=1}^{p} \phi_j \mathrm{E}(X_{t-j} X_{t-k}) \Rightarrow c(k) - \sum_{j=1}^{p} \phi_j c(k-j) = 0. \tag{6.28}$$

Let us now consider the case that the roots of the characteristic polynomial lie both outside and inside the unit circle, thus X_t does not have a causal solution but it is still strictly and second order stationary (with autocovariance, say $\{c(k)\}$). In the previous section we showed that there exists a causal AR(p) $\widetilde{\phi}(B)\widetilde{X}_t = \varepsilon_t$ (where $\phi(B)$ and $\widetilde{\phi}(B) = 1 - \sum_{j=1}^{p} \tilde{\phi}_j z^j$ are the characteristic polynomials defined in (6.24) and (6.26)). We showed that both have the same autocovariance structure. Therefore,

$$c(k) - \sum_{j=1}^{p} \tilde{\phi}_j c(k-j) = 0$$

This means the Yule-Walker equations for $\{X_t\}$ would actually give the AR(p) coefficients of $\{\tilde{X}_t\}$. Thus if the Yule-Walker equations were used to estimate the AR coefficients of $\{X_t\}$, in reality we would be estimating the AR coefficients of the corresponding causal $\{\tilde{X}_t\}$.

6.4.2 Filtering non-causal AR models

Here we discuss the surprising result that filtering a non-causal time series with the corresponding causal AR parameters leaves a sequence which is uncorrelated but not independent. Let us suppose that

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t,$$

where ε_t are iid, $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{var}(\varepsilon_t) < \infty$. It is clear that given the input X_t, if we apply the filter $X_t - \sum_{j=1}^{p} \phi_j X_{t-j}$ we obtain an iid sequence (which is $\{\varepsilon_t\}$).

Suppose that we filter $\{X_t\}$ with the causal coefficients $\{\widetilde{\phi}_j\}$, the output $\widetilde{\varepsilon}_t = X_t - \sum_{j=1}^{p} \widetilde{\phi}_j X_{t-j}$ is not an independent sequence. However, it is an *uncorrelated sequence.* We illustrate this with an example.

Example 6.4.1 *Let us return to the AR(1) example, where $X_t = \phi X_{t-1} + \varepsilon_t$. Let us suppose that $\phi > 1$, which corresponds to a non-causal time series, then X_t has the solution*

$$X_t = -\sum_{j=1}^{\infty} \frac{1}{\phi^j} \varepsilon_{t+j+1}.$$

The causal time series with the same covariance structure as X_t is $\widetilde{X}_t = \frac{1}{\phi}\widetilde{X}_{t-1} + \varepsilon$ (which has backshift representation $(1 - 1/(\phi B))X_t = \varepsilon_t$). Suppose we pass X_t through the causal filter

$$\begin{aligned}\widetilde{\varepsilon}_t &= (1 - \frac{1}{\phi}B)X_t = X_t - \frac{1}{\phi}X_{t-1} = -\frac{(1 - \frac{1}{\phi}B)}{B(1 - \frac{1}{\phi B})}\varepsilon_t \\ &= -\frac{1}{\phi}\varepsilon_t + (1 - \frac{1}{\phi^2})\sum_{j=1}^{\infty} \frac{1}{\phi^{j-1}}\varepsilon_{t+j}.\end{aligned}$$

Evaluating the covariance of the above (assuming wlog that $\mathrm{var}(\varepsilon) = 1$) is

$$\mathrm{cov}(\widetilde{\varepsilon}_t, \widetilde{\varepsilon}_{t+r}) = -\frac{1}{\phi}(1 - \frac{1}{\phi^2})\frac{1}{\phi^r} + (1 - \frac{1}{\phi^2})^2 \sum_{j=0}^{\infty} \frac{1}{\phi^{2j}} = 0.$$

Thus we see that $\{\widetilde{\varepsilon}_t\}$ is an uncorrelated sequence, but unless it is Gaussian it is clearly not independent. One method to study the higher order dependence of $\{\widetilde{\varepsilon}_t\}$, by considering it's higher order cumulant structure etc.

The above above result can be generalised to general AR models, and it is relatively straightforward to prove using the Crámer representation of a stationary process (see Section 10.5, Theorem **??**).

Exercise 6.8 *(i) Consider the causal AR(p) process*

$$X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t.$$

Derive a parallel process with the same autocovariance structure but that is non-causal (it should be real).

(ii) Simulate both from the causal process above and the corresponding non-causal process with non-Gaussian innovations (see Section 4.8). Show that they have the same ACF function.

(iii) Find features which allow you to discriminate between the causal and non-causal process.

Chapter 7

Prediction

Prerequisites

- The best linear predictor.
- Difference between best linear predictors and best predictors.

  [Need to explain]
- Some idea of what a basis of a vector space is.

Objectives

- Understand that prediction using a long past can be difficult because a large matrix has to be inverted, thus alternative, recursive method are often used to avoid direct inversion.
- Understand the derivation of the Levinson-Durbin algorithm, and why the coefficient, $\phi_{t,t}$, corresponds to the partial correlation between X_1 and X_{t+1}.
- Understand how these predictive schemes can be used write space of $\overline{sp}(X_t, X_{t-1}, \ldots, X_1)$ in terms of an orthogonal basis $\overline{sp}(X_t - P_{X_{t-1},X_{t-2},\ldots,X_1}(X_t), \ldots, X_1)$.
- Understand how the above leads to the Wold decomposition of a second order stationary time series.
- To understand how to approximate the prediction for an ARMA time series into a scheme which explicitly uses the ARMA structure. And this approximation improves geometrically, when the past is large.

One motivation behind fitting models to a time series is to forecast future unobserved observations - which would not be possible without a model. In this chapter we consider forecasting, based on the assumption that the model and/or autocovariance structure is known.

7.1 Using prediction in estimation

There are various reasons prediction is important. The first is that forecasting has a vast number of applications from finance to climatology. The second reason is that it forms the basis of most estimation schemes. To understand why forecasting is important in the latter, we now obtain the "likelihood" of the observed time series $\{X_t\}_{t=1}^{n}$. We assume the joint density of $\underline{X}_n = (X_1, \ldots, X_n)$ is $f_n(\underline{x}_n; \theta)$. By using conditioning it is clear that the likelihood is

$$f_n(\underline{x}_n; \theta) = f_1(x_1; \theta) f_2(x_2|x_1; \theta) f_3(x_3|x_2, x_1; \theta) \ldots f_n(x_n|x_{n-1}, \ldots, x_1; \theta)$$

Therefore the log-likelihood is

$$\log f_n(\underline{x}_n; \theta) = \log f_1(x_1) + \sum_{t=1}^{n} \log f_t(x_t|x_{t-1}, \ldots, x_1; \theta).$$

The parameters may be the AR, ARMA, ARCH, GARCH etc parameters. However, usually the conditional distributions $f_t(x_t|x_{t-1}, \ldots, x_1; \theta)$ which make up the joint density $f(\underline{x}; \theta)$ is completely unknown. However, often we can get away with assuming that the conditional distribution is Gaussian and we can still consistently estimate the parameters so long as the model has been correctly specified. Now, if we can "pretend" that the conditional distribution is Gaussian, then all we need is the conditional mean and the conditional variance

$$E(X_t|X_{t-1}, \ldots, X_1; \theta) = \mathrm{E}\left(X_t|X_{t-1}, \ldots, X_1; \theta\right) \text{ and } V(X_t|X_{t-1}, \ldots, X_1, \theta) = \mathrm{var}\left(X_t|X_{t-1}, \ldots, X_1; \theta\right).$$

Using this above and the "Gaussianity" of the conditional distribution gives

$$\log f_t(x_t|x_{t-1}, \ldots, x_1; \theta) = -\frac{1}{2} \log V(x_t|x_{t-1}, \ldots, x_1, \theta) - \frac{(x_t - E(x_t|x_{t-1}, \ldots, x_1, \theta))^2}{V(x_t|x_{t-1}, \ldots, x_1, \theta)}.$$

Using the above the log density

$$\log f_n(\underline{x}_n;\theta) = -\frac{1}{2}\sum_{t=1}^{n}\left(\log V(x_t|x_{t-1},\ldots,x_1,\theta) + \frac{(x_t - E(x_t|x_{t-1},\ldots,x_1,\theta))^2}{V(x_t|x_{t-1},\ldots,x_1,\theta)}\right).$$

Thus the log-likelihood

$$\mathcal{L}(\underline{X}_n;\theta) = -\frac{1}{2}\sum_{t=1}^{n}\left(\log V(X_t|X_{t-1},\ldots,X_1,\theta) + \frac{(X_t - E(X_t|X_{t-1},\ldots,X_1,\theta))^2}{V(X_t|X_{t-1},\ldots,X_1,\theta)}\right).$$

Therefore we observe that in order to evaluate the log-likelihood, and estimate the parameters, we require the conditonal mean and the conditional variance

$$E(X_t|X_{t-1},\ldots,X_1;\theta) \qquad \text{and} \qquad V(X_t|X_{t-1},\ldots,X_1;\theta).$$

This means that in order to do any form of estimation we need a clear understanding of what the conditional mean (which is simply the best predictor of the observation tomorrow given the past) and the conditional variance is for various models.

Note:

- Often expressions for conditional mean and variance can be extremely unwieldy. Therefore, often we require approximations of the conditonal mean and variance which are tractable (this is reminiscent of the Box-Jenkins approach and is till used when the conditional expectation and variance are difficult to estimate).

- Suppose we "pretend" that the time series $\{X_t\}$ is Gaussian. Which we can if it is linear, even if it is not. But we *cannot* if the time series is nonlinear (since nonlinear time series are not Gaussian), then the conditional variance $\text{var}(X_t|X_{t-1},\ldots,X_1)$ will *not* be random (this is a well known result for Gaussian random variables). If X_t is nonlinear, it can be conditionally Gaussian but not Gaussian.

- If the model is linear usually the conditonal expectation $E(X_t|X_{t-1},\ldots,X_1;\theta)$ is replaced with the best linear predictor of X_t given $X_{t-1},\ldots,X_1$. This means if the model is in fact non-causal the estimator will give a causal solution instead. Though not critical it is worth bearing in mind.

7.2 Forecasting for autoregressive processes

Worked example: AR(1) Let

$$X_{t+1} = \phi X_t + \varepsilon_{t+1}$$

where $\{\varepsilon_t\}_t$ are iid random variable. We will assume the process is causal, thus $|\phi| < 1$. Since $\{X_t\}$ are iid random variables, X_{t-1} contains no information about ε_t. Therefore the best linear (indeed best predictor) of X_{t+1} given all the past information is contained in X_t

$$X_t(1) = \phi X_t.$$

To quantify the error in the prediction we use the mean squared error

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

$X_t(1)$ gives the one-step ahead prediction. Since

$$X_{t+2} = \phi X_{t+1} + \varepsilon_{t+2} = \phi^2 X_t + \phi\varepsilon_{t+1} + \varepsilon_{t+2}$$

and $\{\varepsilon_t\}$ are iid random variables, then the best linear predictor (and best predictor) of X_{t+2} given X_t is

$$X_t(2) = \phi X_t(1) = \phi^2 X_{t+1}.$$

Observe it recurses on the previous best linear predictor which makes it very easy to evaluate. The mean squared error in the forecast is

$$\mathrm{E}[X_{t+3} - X_t(2)]^2 = \mathrm{E}[\phi\varepsilon_{t+1} + \varepsilon_{t+2}]^2 = (1 + \phi^2)\mathrm{var}[\varepsilon_t].$$

Using a similar strategy we can forecast r steps into the future:

$$X_t(r) = \phi X_t(r-1) = \phi^r X_t$$

where the mean squared error is

$$\mathrm{E}[X_{t+r} - X_t(r)]^2 = \mathrm{E}[\sum_{i=0}^{r-1} \phi^i \varepsilon_{t+r-i}]^2 = \mathrm{var}[\varepsilon_t] \sum_{i=0}^{r-1} \phi^{2i}.$$

<u>Worked example: AR(2)</u> We now extend the above prediction strategy to AR(2) models (it is straightfoward to go to the AR(p) model). It is best understood using the vector AR representation of the model. Let

$$X_{t+1} = \phi_1 X_t + \phi_2 X_{t-1} + \varepsilon_{t+1}$$

where $\{\varepsilon_t\}_t$ are iid random variables and the characteristic function is causal. We can rewrite the AR(2) as a VAR(1)

$$\begin{aligned} \begin{pmatrix} X_{t+1} \\ X_t \end{pmatrix} &= \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{t+1} \\ 0 \end{pmatrix} \\ \Rightarrow \underline{X}_{t+1} &= \Phi \underline{X}_t + \underline{\varepsilon}_{t+1}. \end{aligned}$$

This looks like a AR(1) and motivates how to forecast into the future. Since ε_{t+1} is independent of $\{X_{t-j}\}_{j \geq 0}$ the best linear predictor of X_{t+1} can be obtained using

$$X_t(1) = \begin{pmatrix} X_{t+1} \\ X_t \end{pmatrix}_{(1)} = \left[\begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} \right]_{(1)}.$$

The mean squared error is $\mathrm{E}[\widehat{X}_t(1) - X_{t+1}]^2 = \sigma^2$. To forecast two steps into the future we use that

$$\underline{X}_{t+2} = \Phi^2 \underline{X}_t + \Phi \underline{\varepsilon}_{t+1} + \underline{\varepsilon}_{t+2}.$$

Thus the best linear predictor of X_{t+2} is

$$X_t(2) = [\Phi^2 \underline{X}_t]_{(1)} = \phi_1(2) X_t + \phi_2(2) X_{t-1},$$

where $[\cdot]_{(1)}$ denotes the first entry in the vector and $(\phi_1(2), \phi_2(2))$ is the first row vector in the matrix Φ^2. The mean squared error is a

$$\mathrm{E}\left(\phi_1\varepsilon_{t+1}+\varepsilon_{t+2}\right)^2=(1+\phi_1^2)\mathrm{var}(\varepsilon_t).$$

We continue this iteration to obtain the r-step ahead predictor

$$X_t(r)=[\Phi\underline{X}_t(r-1)]_{(1)}=[\Phi^r\underline{X}_t]_{(1)}=\phi_1(r)X_t+\phi_2(r)X_{t-1},$$

as above $(\phi_1(r),\phi_2(r))$ is the first row vector in the matrix Φ^r. The mean squared error is

$$\begin{aligned}\mathrm{E}\left(X_{t+r}-X_t(r)\right)^2 &= \mathrm{E}\left(\sum_{i=0}^{r-1}[\Phi^i]_{(1,1)}\varepsilon_{t+r-i}\right)^2\\ &= \mathrm{var}[\varepsilon_t]\sum_{i=0}^{r-1}([\Phi^i]_{(1,1)})^2.\end{aligned}$$

7.3 Forecasting for AR(p)

The above iteration for calculating the best linear predictor easily generalises for any AR(p) process. Let

$$X_{t+1}=\phi_1X_t+\phi_2X_{t-1}+\ldots+\phi_pX_{t+1-p}+\varepsilon_{t+1}$$

where $\{\varepsilon_t\}_t$ are iid random variables and the characteristic function is causal. We can rewrite the AR(p) as a VAR(1)

$$\begin{aligned}\begin{pmatrix}X_{t+1}\\ X_t\\ \vdots\\ \vdots\\ X_{t-p+1}\end{pmatrix} &= \begin{pmatrix}\phi_1 & \phi_2 & \phi_3 & \ldots & \phi_p\\ 1 & 0 & 0 & \ldots & 0\\ 0 & 1 & 0 & \ldots & 0\\ \vdots & \ldots & \vdots & \ddots & 0\\ 0 & 0 & 0 & \ldots & 0\end{pmatrix}\begin{pmatrix}X_t\\ X_{t-1}\\ \vdots\\ \vdots\\ X_{t-p}\end{pmatrix}+\begin{pmatrix}\varepsilon_{t+1}\\ 0\\ 0\\ \vdots\\ 0\end{pmatrix}\\ \Rightarrow\underline{X}_{t+1} &= \Phi\underline{X}_t+\underline{\varepsilon}_{t+1}.\end{aligned}$$

Therefore the r step ahead predictor is

$$X_t(r) = [\Phi \underline{X}_t(r-1)]_{(1)} = [\Phi^r \underline{X}_t]_{(1)} = \sum_{j=1}^{p} \phi_j(r) X_{t+1-j}$$

as above $(\phi_1(r), \phi_2(r), \ldots, \phi_p(r))$ is the first row vector in the matrix Φ^r. The mean squared error is

$$\begin{aligned}
\mathrm{E}\left(X_{t+r} - X_t(r)\right)^2 &= \mathrm{E}\left(\sum_{i=0}^{r-1} [\Phi^i]_{(1,1)} \varepsilon_{t+r-i}\right)^2 \\
&= \mathrm{var}[\varepsilon_t] \sum_{i=0}^{r-1} ([\Phi^i]_{(1,1)})^2 \\
&= \mathrm{var}[\varepsilon_t] \sum_{i=0}^{r-1} \phi_1(i)^2.
\end{aligned}$$

The above predictors are easily obtained using a recursion. However, we now link $\{\phi_j(r)\}_{j=1}^{p}$ to the underlying AR (and MA) coefficients.

Lemma 7.3.1 *Suppose X_t has a causal AR(p) representation*

$$X_{t+1} = \phi_1 X_t + \phi_2 X_{t-1} + \ldots + \phi_p X_{t+1-p} + \varepsilon_{t+1}$$

and

$$X_{t+1} = (1 - \sum_{j=1}^{p} \phi_j B^j) \varepsilon_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

is its MA(∞) representation. Then the predictive coefficients are

$$\phi_j(r) = \sum_{s=0}^{p-j} \phi_{j+s} \psi_{r-1-s} = \sum_{u=0}^{\min(p,j-1)} \phi_u \psi_{r-1+j-u} \qquad r \geq 1$$

and the best r-ahead predictor is

$$X_t(r) = \sum_{j=1}^{p} X_{t+1-j} \sum_{s=0}^{p-j} \phi_{j+s} \psi_{r-1-s} \qquad r \geq 1.$$

The mean squared error is

$$\mathrm{E}[X_{t+r} - X_t(r)]^2 = \mathrm{var}[\varepsilon_t] \sum_{i=0}^{r-1} \psi_i^2$$

with $\psi_0 = 1$*,*

7.4 Forecasting for general time series using infinite past

In the previous section we focussed on time series which had an AR(p) representation. We now consider general time series models and best linear predictors (linear forecasts) for such time series. Specifically, we focus predicting the future given the (unrealistic situation) of the infinite past. Of course, this is an idealized setting, and in the next section we consider linear forecasts based on the finite past (for general stationary time series). A technical assumption we will use in this section is that the stationary time series $\{X_t\}$ has both an AR(∞) and MA(∞) representation (its spectral density bounded away from zero and is finite):

$$X_{t+1} = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t+1-j} = \sum_{j=1}^{\infty} a_j X_{t+1-j} + \varepsilon_{t+1}, \tag{7.1}$$

where $\{\varepsilon_t\}$ are iid random variables (recall Definition 4.5.2). A technical point is that the assumption on $\{\varepsilon_t\}$ can be relaxed to uncorrelated random variables if we are willing to consider best linear predictor and not best predictors. Using (7.2), it is clear the best linear one-ahead predictor is

$$X_t(1) = \sum_{j=1}^{\infty} a_j X_{t+1-j}. \tag{7.2}$$

and the mean squared error is $\mathrm{E}[X_{t+1} - X_t(1)]^2 = \sigma^2$. Transfering the ideas for the AR(p) model (predicting r steps ahead), the best linear predictor r-steps ahead for the general time series is

$$X_t(r) = \sum_{j=1}^{\infty} \phi_j(r) X_{t+1-j} \qquad r \geq 1. \tag{7.3}$$

But analogous to Lemma 7.3.1 we can show that

$$\phi_j(r) = \sum_{s=0}^{\infty} a_{j+s}\psi_{r-1-s} \qquad r \geq 1.$$

Substituting this into (7.3) gives

$$X_t(r) = \sum_{j=1}^{\infty} X_{t+1-j} \sum_{s=0}^{\infty} a_{j+s}\psi_{r-1-s} \qquad r \geq 1.$$

This is not a particularly simple method for estimating the predictors as one goes further in the future. Later in this section we derive a recursion for prediction. First, we obtain the mean squared error in the prediction.

To obtain the mean squared error, we note that since $X_t, X_{t-1}, X_{t-2}, \ldots$ is observed, we can obtain ε_τ (for $\tau \leq t$) by using the invertibility condition

$$\varepsilon_\tau = X_\tau - \sum_{j=1}^{\infty} a_j X_{\tau-i}.$$

This means that given the time series $\{X_{t-j}\}_{j=0}^{\infty}$ (and AR(∞) parameters $\{a_j\}$) we can obtain all the innovations $\{\varepsilon_{t-j}\}_{j=0}^{\infty}$ and visa versa. Based on this we revisit the problem of predicting X_{t+k} given $\{X_\tau; \tau \leq t\}$ but this time in terms of the innovations. Using the MA(∞) presentation (since the time series is causal) of X_{t+k} we have

$$X_{t+r} = \underbrace{\sum_{j=0}^{\infty} \psi_{j+r}\varepsilon_{t-j}}_{\text{innovations are `observed'}} + \underbrace{\sum_{j=0}^{r-1} \psi_j \varepsilon_{t+r-j}}_{\text{future innovations impossible to predict}}.$$

Thus we can write the best predictor of X_{t+r} given $\{X_{t-j}\}_{j=0}^{\infty}$ as

$$\begin{aligned} X_t(r) &= \sum_{j=0}^{\infty} \psi_{j+r}\varepsilon_{t-j} \qquad (7.4) \\ &= \sum_{j=0}^{\infty} \psi_{j+r}\left(X_{t-j} - \sum_{i=1}^{\infty} a_i X_{t-j-i}\right) \\ &= \sum_{j=0}^{\infty} \phi_j(r) X_{t-j}. \end{aligned}$$

Using the above we see that the mean squared error is

$$\mathrm{E}[X_{t+r} - X_t(r)]^2 = \mathrm{E}[\sum_{j=0}^{r-1} \psi_j \varepsilon_{t+r-j}]^2 = \sigma^2 \sum_{j=0}^{r-1} \psi_j^2.$$

We now show how $X_t(r)$ can be evaluated recursively using the invertibility assumption.

Step 1 We use invertibility in (7.2) to give

$$X_t(1) = \sum_{i=1}^{\infty} a_i X_{t+1-i},$$

and $\mathrm{E}\left[X_{t+1} - X_t(1)\right]^2 = \mathrm{var}[\varepsilon_t]$

Step 2 To obtain the 2-step ahead predictor we note that

$$\begin{aligned} X_{t+2} &= \sum_{i=2}^{\infty} a_i X_{t+2-i} + a_1 X_{t+1} + \varepsilon_{t+2} \\ &= \sum_{i=2}^{\infty} a_i X_{t+2-i} + a_1 [X_t(1) + \varepsilon_{t+1}] + \varepsilon_{t+2}, \end{aligned}$$

thus it is clear that

$$X_t(2) = \sum_{i=2}^{\infty} a_i X_{t+2-i} + a_1 X_t(1)$$

and $\mathrm{E}\left[X_{t+2} - X_t(2)\right]^2 = \mathrm{var}[\varepsilon_t]\left(a_1^2 + 1\right) = \mathrm{var}[\varepsilon_t]\left(1 + \psi_1^2\right)$.

Step 3 To obtain the 3-step ahead predictor we note that

$$\begin{aligned} X_{t+3} &= \sum_{i=3}^{\infty} a_i X_{t+2-i} + a_2 X_{t+1} + a_1 X_{t+2} + \varepsilon_{t+3} \\ &= \sum_{i=3}^{\infty} a_i X_{t+2-i} + a_2\left(X_t(1) + \varepsilon_{t+1}\right) + a_1\left(X_t(2) + a_1\varepsilon_{t+1} + \varepsilon_{t+2}\right) + \varepsilon_{t+3}. \end{aligned}$$

Thus

$$X_t(3) = \sum_{i=3}^{\infty} a_i X_{t+2-i} + a_2 X_t(1) + a_1 X_t(2)$$

and $\mathrm{E}\left[X_{t+3} - X_t(3)\right]^2 = \mathrm{var}[\varepsilon_t]\left[(a_2 + a_1^2)^2 + a_1^2 + 1\right] = \mathrm{var}[\varepsilon_t]\left(1 + \psi_1^2 + \psi_2^2\right)$.

Step r Using the arguments it can be shown that

$$X_t(r) = \underbrace{\sum_{i=r}^{\infty} a_i X_{t+r-i}}_{\text{observed}} + \sum_{i=1}^{r-1} a_i \underbrace{X_t(r-i)}_{\text{predicted}}.$$

And we have already shown that $\mathrm{E}[X_{t+r} - X_t(r)]^2 = \sigma^2 \sum_{j=0}^{r-1} \psi_j^2$

Thus the r-step ahead predictor can be recursively estimated.

We note that the predictor given above is based on the assumption that the infinite past is observed. In practice this is not a realistic assumption. However, in the special case that time series is an autoregressive process of order p (with AR parameters $\{\phi_j\}_{j=1}^{p}$) and $X_t, \ldots, X_{t-m}$ is observed where $m \geq p-1$, then the above scheme can be used for forecasting. More precisely,

$$\begin{aligned}
X_t(1) &= \sum_{j=1}^{p} \phi_j X_{t+1-j} \\
X_t(r) &= \sum_{j=r}^{p} \phi_j X_{t+r-j} + \sum_{j=1}^{r-1} \phi_j X_t(r-j) \text{ for } 2 \leq r \leq p \\
X_t(r) &= \sum_{j=1}^{p} \phi_j X_t(r-j) \text{ for } r > p. \qquad (7.5)
\end{aligned}$$

However, in the general case more sophisticated algorithms are required when only the finite past is known.

7.4.1 Example: Forecasting yearly temperatures

We now fit an autoregressive model to the yearly temperatures from 1880-2008 and use this model to forecast the temperatures from 2009-2013. In Figure 7.1 we give a plot of the temperature time series together with its ACF. It is clear there is some trend in the temperature data, therefore we have taken second differences, a plot of the second difference and its ACF is given in Figure 7.2. We now use the command `ar.yule(res1,order.max=10)` (we will discuss in Chapter 9 how this function estimates the AR parameters) to estimate the the AR parameters.

Remark 7.4.1 (The Yule-Walker estimator in prediction) *The least squares estimator (or equivalently the conditional likelihood) is likely to give a causal estimator of the AR parameters. But it is not guaranteed. On the other hand the Yule-Walker estimator is guaranteed to give a causal*

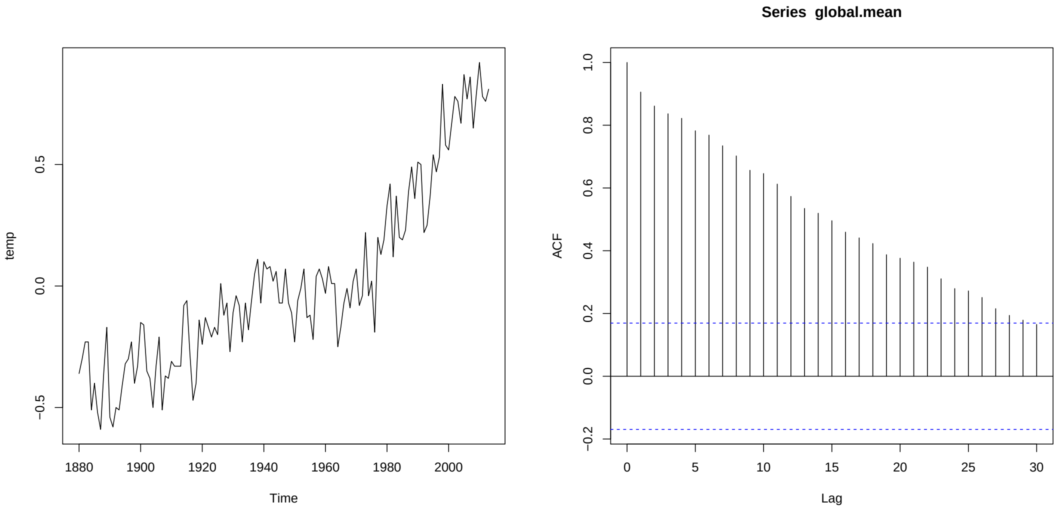


Figure 7.1: Yearly temperature from 1880-2013 and the ACF.

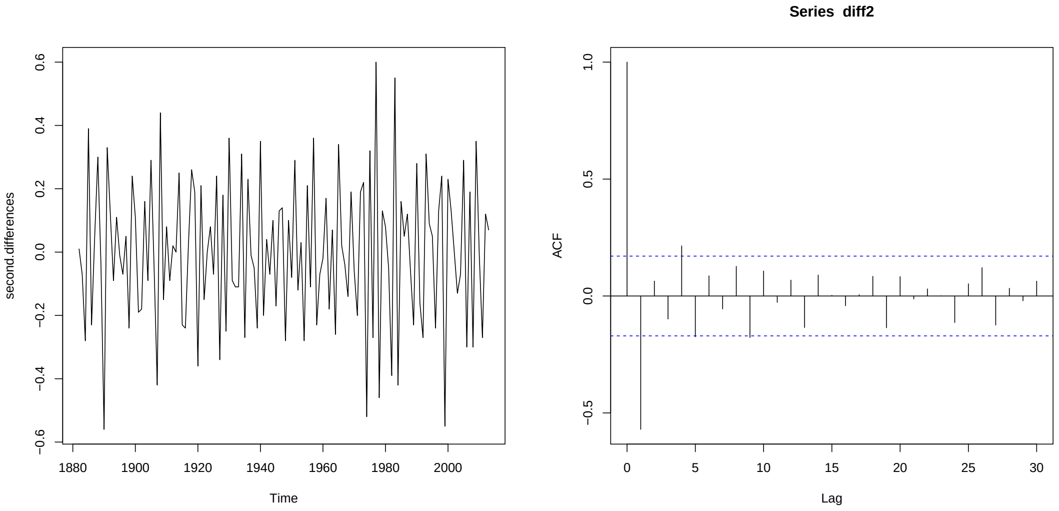


Figure 7.2: Second differences of yearly temperature from 1880-2013 and its ACF.

solution. This will matter for prediction. We emphasize here that the least squares estimator cannot consistently estimate non-causal solutions, it is only a quirk of the estimation method that means at times the solution may be noncausal.

If the time series $\{X_t\}_t$ is linear and stationary with mean zero, then if we predict several steps into the future we would expect our predictor to be close to zero (since $\mathrm{E}(X_t) = 0$). This is guaranteed if one uses AR parameters which are causal (since the eigenvalues of the VAR matrix is less than one); such as the Yule-Walker estimators. On the other hand, if the parameter estimators do *not correspond to a causal solution (as could happen for the least squares estimator), the predictors may explode for long term forecasts which makes no sense.*

The function `ar.yule` uses the AIC to select the order of the AR model. When fitting the second differences from (from 1880-2008 - a data set of length of 127) the AIC chooses the AR(7) model

$$X_t = -1.1472X_{t-1} - 1.1565X_{t-2} - 1.0784X_{t-3} - 0.7745X_{t-4} - 0.6132X_{t-5} - 0.3515X_{t-6} - 0.1575X_{t-7} + \varepsilon_t,$$

with $\text{var}[\varepsilon_t] = \sigma^2 = 0.02294$. An ACF plot after fitting this model and then estimating the residuals $\{\varepsilon_t\}$ is given in Figure 7.3. We observe that the ACF of the residuals 'appears' to be uncorrelated, which suggests that the AR(7) model fitted the data well. Later we define the Ljung-Box test, which is a method for checking this claim. However since the residuals are *estimated* residuals and *not* the true residual, the results of this test need to be taken with a large pinch of salt. We will show that when the residuals are estimated from the data the error bars given in the ACF plot are not correct and the Ljung-Box test is not pivotal (as is assumed when deriving the limiting distribution under the null the model is correct). By using the sequence of equations

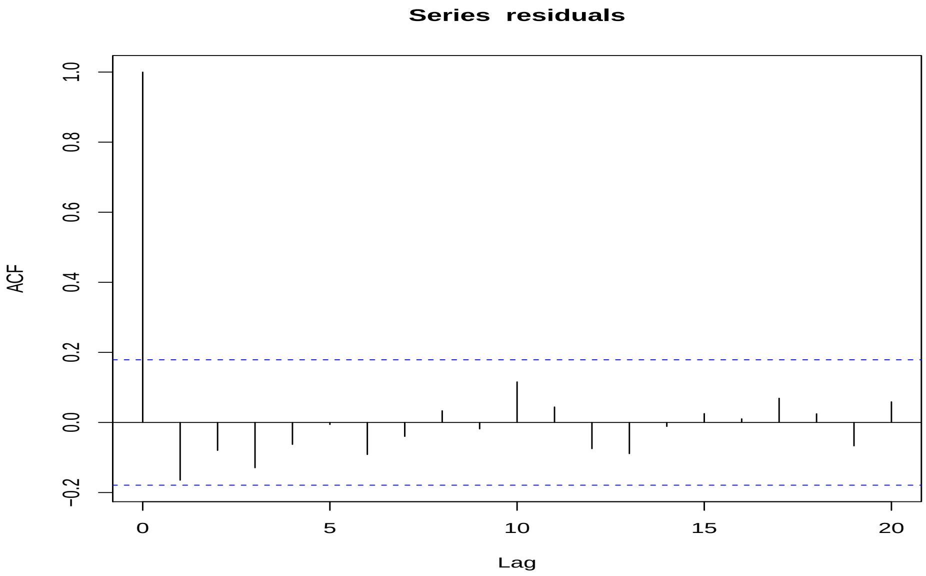


Figure 7.3: An ACF plot of the estimated residuals $\{\widehat{\varepsilon}_t\}$.

$$\begin{aligned}
\hat{X}_{127}(1) &= -1.1472X_{127} - 1.1565X_{126} - 1.0784X_{125} - 0.7745X_{124} - 0.6132X_{123} \\
&\quad -0.3515X_{122} - 0.1575X_{121} \\
\hat{X}_{127}(2) &= -1.1472\hat{X}_{127}(1) - 1.1565X_{127} - 1.0784X_{126} - 0.7745X_{125} - 0.6132X_{124} \\
&\quad -0.3515X_{123} - 0.1575X_{122} \\
\hat{X}_{127}(3) &= -1.1472\hat{X}_{127}(2) - 1.1565\hat{X}_{127}(1) - 1.0784X_{127} - 0.7745X_{126} - 0.6132X_{125} \\
&\quad -0.3515X_{124} - 0.1575X_{123} \\
\hat{X}_{127}(4) &= -1.1472\hat{X}_{127}(3) - 1.1565\hat{X}_{127}(2) - 1.0784\hat{X}_{127}(1) - 0.7745X_{127} - 0.6132X_{126} \\
&\quad -0.3515X_{125} - 0.1575X_{124} \\
\hat{X}_{127}(5) &= -1.1472\hat{X}_{127}(4) - 1.1565\hat{X}_{127}(3) - 1.0784\hat{X}_{127}(2) - 0.7745\hat{X}_{127}(1) - 0.6132X_{127} \\
&\quad -0.3515X_{126} - 0.1575X_{125}.
\end{aligned}$$

We can use $\hat{X}_{127}(1), \ldots, \hat{X}_{127}(5)$ as forecasts of $X_{128}, \ldots, X_{132}$ (we recall are the second differences), which we then use to construct forecasts of the temperatures. A plot of the second difference forecasts together with the true values are given in Figure 7.4. From the forecasts of the second differences we can obtain forecasts of the original data. Let Y_t denote the temperature at time t and X_t its second difference. Then $Y_t = -Y_{t-2} + 2Y_{t-1} + X_t$. Using this we have

$$\begin{aligned}
\widehat{Y}_{127}(1) &= -Y_{126} + 2Y_{127} + X_{127}(1) \\
\widehat{Y}_{127}(2) &= -Y_{127} + 2Y_{127}(1) + X_{127}(2) \\
\widehat{Y}_{127}(3) &= -Y_{127}(1) + 2Y_{127}(2) + X_{127}(3)
\end{aligned}$$

and so forth.

We note that (**??**) can be used to give the mse error. For example

$$\begin{aligned}
\mathrm{E}[X_{128} - \hat{X}_{127}(1)]^2 &= \sigma_t^2 \\
\mathrm{E}[X_{128} - \hat{X}_{127}(1)]^2 &= (1+\phi_1^2)\sigma_t^2
\end{aligned}$$

If we believe the residuals are Gaussian we can use the mean squared error to construct confidence intervals for the predictions. Assuming for now that the parameter estimates are the true parameters (this is not the case), and $X_t = \sum_{j=0}^{\infty} \psi_j(\widehat{\phi})\varepsilon_{t-j}$ is the MA(∞) representation of the AR(7) model, the mean square error for the kth ahead predictor is

$$\sigma^2 \sum_{j=0}^{k-1} \psi_j(\widehat{\phi})^2 \text{ (using (\textbf{??}))}$$

thus the 95% CI for the prediction is

$$\left[X_t(k) \pm 1.96\sigma^2 \sum_{j=0}^{k-1} \psi_j(\widehat{\phi})^2 \right],$$

however this confidence interval for not take into account $X_t(k)$ uses only parameter estimators and not the true values. In reality we need to take into account the approximation error here too.

If the residuals are not Gaussian, the above interval is not a 95% confidence interval for the prediction. One way to account for the non-Gaussianity is to use bootstrap. Specifically, we rewrite the AR(7) process as an MA(∞) process

$$X_t = \sum_{j=0}^{\infty} \psi_j(\widehat{\phi})\varepsilon_{t-j}.$$

Hence the best linear predictor can be rewritten as

$$X_t(k) = \sum_{j=k}^{\infty} \psi_j(\widehat{\phi})\varepsilon_{t+k-j}$$

thus giving the prediction error

$$X_{t+k} - X_t(k) = \sum_{j=0}^{k-1} \psi_j(\widehat{\phi})\varepsilon_{t+k-j}.$$

We have the prediction estimates, therefore all we need is to obtain the distribution of $\sum_{j=0}^{k-1} \psi_j(\widehat{\phi})\varepsilon_{t+k-j}$. This can be done by estimating the residuals and then using bootstrap[1] to estimate the distribution of $\sum_{j=0}^{k-1} \psi_j(\widehat{\phi})\varepsilon_{t+k-j}$, using the empirical distribution of $\sum_{j=0}^{k-1} \psi_j(\widehat{\phi})\varepsilon^*_{t+k-j}$. From this we can

[1] Residual bootstrap is based on sampling from the empirical distribution of the residuals i.e. construct the "bootstrap" sequence $\{\varepsilon^*_{t+k-j}\}_j$ by sampling from the empirical distribution $\widehat{F}(x) = \frac{1}{n}\sum_{t=p+1}^{n} I(\widehat{\varepsilon}_t \leq x)$ (where $\widehat{\varepsilon}_t = X_t - \sum_{j=1}^{p} \widehat{\phi}_j X_{t-j}$). This sequence is used to construct the bootstrap estimator $\sum_{j=0}^{k-1} \psi_j(\widehat{\phi})\varepsilon^*_{t+k-j}$. By doing this several thousand times we can evaluate the empirical distribution of $\sum_{j=0}^{k-1} \psi_j(\widehat{\phi})\varepsilon^*_{t+k-j}$ using these bootstrap samples. This is an estimator of the distribution function of $\sum_{j=0}^{k-1} \psi_j(\widehat{\phi})\varepsilon_{t+k-j}$.

construct the 95% CI for the forecasts.

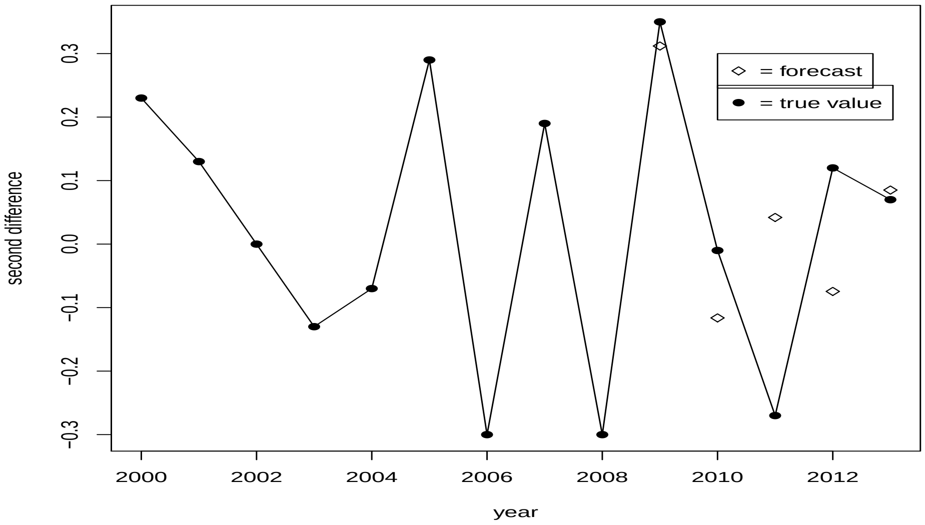


Figure 7.4: Forecasts of second differences.

A small criticism of our approach is that we have fitted a rather large AR(7) model to time series of length of 127. It may be more appropriate to fit an ARMA model to this time series.

Exercise 7.1 *In this exercise we analyze the Sunspot data found on the course website. In the data analysis below only use the data from 1700 - 2003 (the remaining data we will use for prediction). In this section you will need to use the function* `ar.yw` *in* `R`*.*

(i) Fit the following models to the data and study the residuals (using the ACF). Using this decide which model

$$\begin{aligned} X_t &= \mu + A\cos(\omega t) + B\sin(\omega t) + \underbrace{\varepsilon_t}_{AR} \quad \textit{or} \\ X_t &= \mu + \underbrace{\varepsilon_t}_{AR} \end{aligned}$$

is more appropriate (take into account the number of parameters estimated overall).

(ii) Use these models to forecast the sunspot numbers from 2004-2013.

7.5 One-step ahead predictors based on the finite past

We return to Section 6.2.3 and call the definition of the best fitting AR(p) model.

The best fitting AR(p) Let us suppose that $\{X_t\}$ is a general second order stationary time series with autocovariance $\{c(r)\}_r$. We consider the projection of X_t onto $\underline{Y} = (X_{t-p}, \ldots, X_{t-1})$ (technically we should should say $\text{sp}(X_{t-p}, \ldots, X_{t-1})$), this is

$$P_{\underline{Y}}(X_t) = \sum_{j=1}^{p} \phi_{p,j} X_{t-j}$$

where

$$\begin{pmatrix} \phi_{p,1} \\ \vdots \\ \phi_{p.p} \end{pmatrix} = \Sigma_p^{-1} \underline{r}_p, \tag{7.6}$$

where $(\Sigma_p)_{i,j} = c(i-j)$ and $(\underline{r}_p)_i = c(i+1)$. We recall that $X_t - P_{\underline{Y}}(X_t)$ and $\underline{Y}$ are uncorrelated but $X_t - P_{\underline{Y}}(X_t)$ is not necessarily uncorrelated with $\{X_{t-j}\}$ for $j \geq (p+1)$. We call $\{\phi_{p,j}\}$ the best fitting AR(p) coefficients, because if the true model were an AR(p) model $\phi_{p,j} = \phi_j$.

Since $X_t - P_{\underline{Y}}(X_t)$ is uncorrelated with $\underline{Y} = (X_{t-p}, \ldots, X_{t-1})$, the best linear predictor of X_t given $Y = (X_{t-p}, \ldots, X_{t-1})$ is

$$P_{\underline{Y}}(X_t) = \sum_{j=1}^{p} \phi_{p,j} X_{t-j}$$

7.5.1 Levinson-Durbin algorithm

The Levinson-Durbin algorithm, which we describe below forms the basis of several estimation algorithms for linear time series. These include (a) the Gaussian Maximum likelihood estimator, (b) the Yule-Walker estimator and (c) the Burg algorithm. We describe these methods in Chapter 9. But we start with a description of the Levinson-Durbin algorithm.

The Levinson-Durbin algorithm is a method for evaluating $\{\phi_{p,j}\}_{j=1}^{p}$ for an increasing number of past regressors (under the assumption of second order stationarity). A brute force method is to evaluate $\{\phi_{p,j}\}_{j=1}^{p}$ using (7.15), where Σ_p^{-1} is evaluated using standard methods, such as Gauss-Jordan elimination. To solve this system of equations requires $O(p^3)$ operations. The beauty of the Levinson-Durbin algorithm is that it exploits the (Toeplitz) structure of Σ_p to reduce the number of operations to $O(p^2)$. It is evaluated recursively by increasing the order of lags p. It was first proposed in the 1940s by Norman Levinson (for Toeplitz equations). In the 1960s, Jim Durbin adapted the algorithm to time series and improved it. In the discussion below we switch p to t.

We recall that in the aim in one-step ahead prediction is to predict X_{t+1} given $X_t, X_{t-1}, \ldots, X_1$. The best linear predictor is

$$X_{t+1|t} = P_{X_1,\ldots,X_t}(X_{t+1}) = X_{t+1|t,\ldots,1} = \sum_{j=1}^{t} \phi_{t,j} X_{t+1-j}. \tag{7.7}$$

The notation can get a little heavy. But the important point to remember is that as t grows we are not predicting further into the future. We are including more of the past in the one-step ahead prediction.

We first outline the algorithm. We recall that the best linear predictor of X_{t+1} given $X_t, \ldots, X_1$ is

$$X_{t+1|t} = \sum_{j=1}^{t} \phi_{t,j} X_{t+1-j}. \tag{7.8}$$

The mean squared error is $r(t+1) = \mathrm{E}[X_{t+1} - X_{t+1|t}]^2$. Given that the second order stationary covariance structure, the idea of the Levinson-Durbin algorithm is to recursively estimate $\{\phi_{t,j}; j = 1, \ldots, t\}$ given $\{\phi_{t-1,j}; j = 1, \ldots, t-1\}$ (which are the coefficients of the best linear predictor of X_t given $X_{t-1}, \ldots, X_1$). Let us suppose that the autocovariance function $c(k) = \mathrm{cov}[X_0, X_k]$ is known. The Levinson-Durbin algorithm is calculated using the following recursion.

Step 1 $\phi_{1,1} = c(1)/c(0)$ and $r(2) = \mathrm{E}[X_2 - X_{2|1}]^2 = \mathrm{E}[X_2 - \phi_{1,1}X_1]^2 = c(0) - \phi_{1,1}c(1)$.

Step 2 For $j = t$

$$\begin{aligned}
\phi_{t,t} &= \frac{c(t) - \sum_{j=1}^{t-1} \phi_{t-1,j} c(t-j)}{r(t)} \\
\phi_{t,j} &= \phi_{t-1,j} - \phi_{t,t}\phi_{t-1,t-j} \qquad 1 \leq j \leq t-1,
\end{aligned}$$

Step 3 $r(t+1) = r(t)(1 - \phi_{t,t}^2)$.

We give two proofs of the above recursion.

Exercise 7.2 *(i) Suppose $X_t = \phi X_{t-1} + \varepsilon_t$ (where $|\phi| < 1$). Use the Levinson-Durbin algorithm, to deduce an expression for $\phi_{t,j}$ for $(1 \leq j \leq t)$.*

(ii) Suppose $X_t = \phi\varepsilon_{t-1} + \varepsilon_t$ (where $|\phi| < 1$). Use the Levinson-Durbin algorithm (and possibly Maple/Matlab), deduce an expression for $\phi_{t,j}$ for $(1 \leq j \leq t)$. (recall from Exercise 6.3 that you already have an analytic expression for $\phi_{t,t}$).

7.5.2 A proof of the Durbin-Levinson algorithm based on projections

Let us suppose $\{X_t\}$ is a zero mean stationary time series and $c(k) = \mathrm{E}(X_k X_0)$. Let $P_{X_t,\ldots,X_2}(X_1)$ denote the best linear predictor of X_1 given $X_t,\ldots,X_2$ and $P_{X_t,\ldots,X_2}(X_{t+1})$ denote the best linear predictor of X_{t+1} given $X_t,\ldots,X_2$. Stationarity means that the following predictors share the same coefficients

$$\begin{aligned} X_{t|t-1} &= \sum_{j=1}^{t-1}\phi_{t-1,j}X_{t-j} \qquad P_{X_t,\ldots,X_2}(X_{t+1}) = \sum_{j=1}^{t-1}\phi_{t-1,j}X_{t+1-j} \qquad (7.9)\\ P_{X_t,\ldots,X_2}(X_1) &= \sum_{j=1}^{t-1}\phi_{t-1,j}X_{j+1}. \end{aligned}$$

The last line is because stationarity means that flipping a time series round has the same correlation structure. These three relations are an important component of the proof.

Recall our objective is to derive the coefficients of the best linear predictor of $P_{X_t,\ldots,X_1}(X_{t+1})$ based on the coefficients of the best linear predictor $P_{X_{t-1},\ldots,X_1}(X_t)$. To do this we partition the space $\overline{\mathrm{sp}}(X_t,\ldots,X_2,X_1)$ into two orthogonal spaces $\overline{\mathrm{sp}}(X_t,\ldots,X_2,X_1) = \overline{\mathrm{sp}}(X_t,\ldots,X_2,X_1) \oplus \overline{\mathrm{sp}}(X_1 - P_{X_t,\ldots,X_2}(X_1))$. Therefore by uncorrelatedness we have the partition

$$\begin{aligned} X_{t+1|t} &= P_{X_t,\ldots,X_2}(X_{t+1}) + P_{X_1 - P_{X_t,\ldots,X_2}(X_1)}(X_{t+1}) \\ &= \underbrace{\sum_{j=1}^{t-1}\phi_{t-1,j}X_{t+1-j}}_{\text{by (7.9)}} + \underbrace{\phi_{tt}\left(X_1 - P_{X_t,\ldots,X_2}(X_1)\right)}_{\text{by projection onto one variable}} \\ &= \sum_{j=1}^{t-1}\phi_{t-1,j}X_{t+1-j} + \phi_{t,t}\left(X_1 - \underbrace{\sum_{j=1}^{t-1}\phi_{t-1,j}X_{j+1}}_{\text{by (7.9)}}\right). \qquad (7.10) \end{aligned}$$

We start by evaluating an expression for $\phi_{t,t}$ (which in turn will give the expression for the other coefficients). It is straightforward to see that

$$\begin{aligned}\phi_{t,t} &= \frac{\mathrm{E}(X_{t+1}(X_1 - P_{X_t,\ldots,X_2}(X_1)))}{\mathrm{E}(X_1 - P_{X_t,\ldots,X_2}(X_1))^2} \\ &= \frac{\mathrm{E}[(X_{t+1} - P_{X_t,\ldots,X_2}(X_{t+1}) + P_{X_t,\ldots,X_2}(X_{t+1}))(X_1 - P_{X_t,\ldots,X_2}(X_1))]}{\mathrm{E}(X_1 - P_{X_t,\ldots,X_2}(X_1))^2} \\ &= \frac{\mathrm{E}[(X_{t+1} - P_{X_t,\ldots,X_2}(X_{t+1}))(X_1 - P_{X_t,\ldots,X_2}(X_1))]}{\mathrm{E}(X_1 - P_{X_t,\ldots,X_2}(X_1))^2}\end{aligned} \tag{7.11}$$

Therefore we see that the numerator of $\phi_{t,t}$ is the partial covariance between X_{t+1} and X_1 (see Section 6.2), furthermore the denominator of $\phi_{t,t}$ is the mean squared prediction error, since by stationarity

$$\mathrm{E}(X_1 - P_{X_t,\ldots,X_2}(X_1))^2 = \mathrm{E}(X_t - P_{X_{t-1},\ldots,X_1}(X_t))^2 = r(t) \tag{7.12}$$

Returning to (7.11), expanding out the expectation in the numerator and using (7.12) we have

$$\phi_{t,t} = \frac{\mathrm{E}(X_{t+1}(X_1 - P_{X_t,\ldots,X_2}(X_1)))}{r(t)} = \frac{c(0) - \mathrm{E}[X_{t+1}P_{X_t,\ldots,X_2}(X_1))]}{r(t)} = \frac{c(0) - \sum_{j=1}^{t-1}\phi_{t-1,j}c(t-j)}{r(t)}, \tag{7.13}$$

which immediately gives us the first equation in Step 2 of the Levinson-Durbin algorithm. To obtain the recursion for $\phi_{t,j}$ we use (7.10) to give

$$\begin{aligned}X_{t+1|t} &= \sum_{j=1}^{t}\phi_{t,j}X_{t+1-j} \\ &= \sum_{j=1}^{t-1}\phi_{t-1,j}X_{t+1-j} + \phi_{t,t}\left(X_1 - \sum_{j=1}^{t-1}\phi_{t-1,j}X_{j+1}\right).\end{aligned}$$

To obtain the recursion we simply compare coefficients to give

$$\phi_{t,j} = \phi_{t-1,j} - \phi_{t,t}\phi_{t-1,t-j} \qquad 1 \leq j \leq t-1.$$

This gives the middle equation in Step 2. To obtain the recursion for the mean squared prediction error we note that by orthogonality of $\{X_t, \ldots, X_2\}$ and $X_1 - P_{X_t,\ldots,X_2}(X_1)$ we use (7.10) to give

$$\begin{aligned} r(t+1) &= \mathrm{E}(X_{t+1} - X_{t+1|t})^2 = \mathrm{E}[X_{t+1} - P_{X_t,\ldots,X_2}(X_{t+1}) - \phi_{t,t}(X_1 - P_{X_t,\ldots,X_2}(X_1)]^2 \\ &= \mathrm{E}[X_{t+1} - P_{X_2,\ldots,X_t}(X_{t+1})]^2 + \phi_{t,t}^2 \mathrm{E}[X_1 - P_{X_t,\ldots,X_2}(X_1)]^2 \\ & \quad -2\phi_{t,t}\mathrm{E}[(X_{t+1} - P_{X_t,\ldots,X_2}(X_{t+1}))(X_1 - P_{X_t,\ldots,X_2}(X_1))] \\ &= r(t) + \phi_{t,t}^2 r(t) - 2\phi_{t,t} \underbrace{\mathrm{E}[X_{t+1}(X_1 - P_{X_t,\ldots,X_2}(X_1))]}_{=r(t)\phi_{t,t} \text{ by (7.13)}} \\ &= r(t)[1 - \phi_{tt}^2]. \end{aligned}$$

This gives the final part of the equation in Step 2 of the Levinson-Durbin algorithm.

7.5.3 Applying the Durbin-Levinson to obtain the Cholesky decomposition

We recall from Section 5.5 that by sequentially projecting the elements of random vector on the past elements in the vector gives rise to Cholesky decomposition of the inverse of the variance/covariance (precision) matrix. This is exactly what was done in when we make the Durbin-Levinson algorithm. In other words,

$$\mathrm{var}\begin{pmatrix} \frac{X_1}{\sqrt{r(1)}} \\ \frac{X_1 - \phi_{1,1}X_2}{\sqrt{r(2)}} \\ \vdots \\ \frac{X_n - \sum_{j=1}^{n-1}\phi_{n-1,j}X_{n-j}}{\sqrt{r(n)}} \end{pmatrix} = I_n$$

Therefore, if $\Sigma_n = \mathrm{var}[\underline{X}_n]$, where $\underline{X}_n = (X_1, \ldots, X_n)$, then $\Sigma_n^{-1} = L_n D_n L_n'$, where

$$L_n = \begin{pmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ -\phi_{1,1} & 1 & 0 & \ldots & \ldots & 0 \\ -\phi_{2,2} & -\phi_{2,1} & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ -\phi_{n-1,n-1} & -\phi_{n-1,n-2} & -\phi_{n-1,n-3} & \ldots & \ldots & 1 \end{pmatrix} \qquad (7.14)$$

and $D_n = \mathrm{diag}(r_1^{-1}, r_2^{-1}, \ldots, r_n^{-1})$.

7.6 Comparing finite and infinite predictors (advanced)

We recall that

$$X_{t+1|t} = P_{X_t,\ldots,X_1}(X_{t+1}) = \sum_{j=1}^{t} \phi_{t,j} X_{t-j},$$

which is the best linear predictor given the finite past. However, often $\phi_{t,j}$ can be difficult to evaluate (usually with the Durbin-Levinson algorithm) in comparison to the AR(∞) parameters. Thus we define the above approximation

$$\widehat{X}_{t+1|t} = \sum_{j=1}^{t} \phi_j X_{t-j}.$$

How good an approximation $\widehat{X}_{t+1|t}$ is of $X_{t+1|t}$ is given by Baxter's inequality.

Theorem 7.6.1 (Baxter's inequality) *Suppose $\{X_t\}$ has an $AR(\infty)$ representation with parameters $\{\phi_j\}_{j=1}^{\infty}$ such that $\sum_{j=1}^{\infty} |\phi_j| < \infty$. Let $\{\phi_{n,j}\}_{j=1}^{n}$ denote the parameters of the parameters of the best linear predictor of X_{t+1} given $\{X_j\}_{j=1}^{t}$. Then if n is large enough we have*

$$\sum_{j=1}^{n} |\phi_{n,j} - \phi_j| \leq C \sum_{j=n+1}^{\infty} |\phi_j|,$$

where C is a constant that depends on the underlying spectral density.

We note that since $\sum_{j=1}^{\infty} |\phi_j| < \infty$, then $\sum_{j=n+1}^{\infty} |\phi_j| \to 0$ as $n \to \infty$. Thus as n gets large

$$\sum_{j=1}^{n} |\phi_{n,j} - \phi_j| \approx 0.$$

We apply this result to measuring the difference between $X_{t+1|t}$ and $\widehat{X}_{t+1|t}$

$$\mathrm{E}|X_{t+1|t} - \widehat{X}_{t+1|t}| \leq \sum_{j=1}^{t} |\phi_{t,j} - \phi_j| \, \mathrm{E}|X_{t-j}| \leq \mathrm{E}|X_{t-j}| \sum_{j=1}^{t} |\phi_{t,j} - \phi_j| \leq C\mathrm{E}|X_t| \sum_{j=t+1}^{\infty} |\phi_j|.$$

Therefore the best linear predictor and its approximation are "close" for large t.

7.7 r-step ahead predictors based on the finite past

Let $\underline{Y} = (X_{t-p}, \ldots, X_{t-1})$

$$P_{\underline{Y}}(X_{t+r}) = \sum_{j=1}^{p} \phi_{p,j}(r) X_{t-j}$$

where

$$\left(\begin{array}{c} \phi_{p,1}(r) \\ \vdots \\ \phi_{p,p}(r) \end{array}\right) = \Sigma_p^{-1} \underline{r}_{p,r}, \qquad (7.15)$$

where $(\Sigma_p)_{i,j} = c(i-j)$ and $(\underline{r}_{p,r})_i = c(i+r)$. This gives the best finite predictor for the time series at lag r. In practice, one often finds the best fitting AR(p) model, which gives the best finite predictor at lag one. And then uses the AR prediction method described in Section 7.3 to predict forward

$$\widehat{X}_t(r) = [\Phi \underline{\widehat{X}}_t(r-1)]_{(1)} = [\Phi_p^r \underline{X}_t]_{(1)} = \sum_{j=1}^{p} \phi_j(r,p) X_{t+1-j}$$

where

$$\Phi_p = \left(\begin{array}{ccccc} \phi_{p,1} & \phi_{p,2} & \phi_3 & \ldots & \phi_{p.p} \\ 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ldots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \ldots & 0 \end{array}\right).$$

If the true model is not an AR(p) this will not give the best linear predictor, but it will given an approximation of it. Suppose that $j > n$

For ARMA models

$$\sum_{j=1}^{n} |\phi_j(\tau;p) - \phi_{j,n}(\tau)||X_{t-j}| = \begin{cases} O_p\left(\rho^p\right) & \tau \leq p \\ O_p\left(\rho^p \rho^{|\tau-p|}\right) & \tau > p. \end{cases}$$

Lemma 7.7.1 *Suppose the MA(∞) and AR(∞) parameters satisfy $\sum_j |j^K \psi_j| < \infty$ and $\sum_j |j^K a_j| <$*

∞ *for some* $K > 1$*. Then*

$$\sum_{j=1}^{n} |\phi_j(\tau;p) - \phi_j(\tau)| \begin{cases} O\left(\frac{1}{p^K}\right) & \tau \leq p \\ O\left(\frac{1}{p^K|\tau-p|^K}\right) & \tau > p. \end{cases}$$

PROOF. If $\tau < p$

$$\sum_{j=1}^{n} |\phi_j(\tau; f_p) - \phi_j(\tau, f)| = \sum_{j=1}^{n} |\sum_{s=1}^{\infty} \phi_{j+s}(f_p)\psi_{\tau-s}(f_p) - \sum_{s=1}^{\infty} \phi_{j+s}(f)\psi_{\tau-s}(f)| = O\left(\frac{1}{p^K}\right).$$

If $\tau > p$

$$\sum_{j=1}^{n} |\phi_j(\tau; f_p) - \phi_j(\tau, f)| = \sum_{j=1}^{n} |\sum_{s=1}^{\infty} \phi_{j+s}(f_p)\psi_{\tau-s}(f_p) - \sum_{s=1}^{\infty} \phi_{j+s}(f)\psi_{\tau-s}(f)| = O\left(\frac{1}{p^K|\tau-p|^K}\right).$$

7.8 Forecasting for ARMA processes

Given the autocovariance of any stationary process the Levinson-Durbin algorithm allows us to systematically obtain one-step predictors of second order stationary time series without directly inverting a matrix. In this section we consider the special case of ARMA(p,q) models where the ARMA coefficients are known.

For AR(p) models prediction is especially easy, if the number of observations in the finite past, t, is such that $p \leq t$. For $1 \leq t \leq p$ one would use the Durbin-Levinson algorithm and for $t > p$ we use

$$X_{t+1|t} = \sum_{j=1}^{p} \phi_j X_{t+1-j}.$$

For ARMA(p,q) models prediction is not so straightforward, but we show below some simple approximations can be made.

We recall that a causal invertible ARMA(p,q) has the representation

$$X_{t+1} = \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i \varepsilon_{t+1-i} + \varepsilon_{t+1}.$$

Then if the infinite past were observed by using equation (7.4) and the AR(∞) and MA(∞) representation of the ARMA model the best linear predictor is

$$
\begin{aligned}
X_t(1) &= \sum_{j=1}^{\infty}\psi_j\varepsilon_{t+1-j} \\
&= \sum_{j=1}^{\infty}a_jX_{t+1-j}
\end{aligned}
$$

where $\{\psi_j\}$ and $\{a_j\}$ are the AR(∞) and MA(∞) coefficients respectively. The above representation does not explictly use the ARMA representation. However since $\varepsilon_{t-j} = X_{t-j} - X_{t-j-1}(1)$ it is easily seen that an alternative representation is

$$X_t(1) = \sum_{j=1}^{p}\phi_jX_{t+1-j} + \sum_{i=1}^{q}\theta_i\left(X_{t+1-i} - X_{t-i}(1)\right).$$

However, for finite predictors the actual one-step ahead prediction formula is not so simple. It can be shown that for $t \geq \max(p,q)$

$$X_{t+1|t} = \sum_{j=1}^{p}\phi_jX_{t+1-j} + \sum_{i=1}^{q}\theta_{t,i}(X_{t+1-i} - X_{t+1-i|t-i}), \tag{7.16}$$

where the coefficients $\theta_{t,i}$ which can be evaluated from the autocovariance structure of the MA process. A proof is given in the appendix. It can be shown that $\theta_{t,i} \to \theta_i$ as $t \to \infty$ (see Brockwell and Davis (1998)), Chapter 5.

The prediction can be simplified if we make a simple approximation (which works well if t is relatively large). For $1 \leq t \leq \max(p,q)$, set $\widehat{X}_{t+1|t} = X_t$ and for $t > \max(p,q)$ we define the recursion

$$\widehat{X}_{t+1|t} = \sum_{j=1}^{p}\phi_jX_{t+1-j} + \sum_{i=1}^{q}\theta_i(X_{t+1-i} - \widehat{X}_{t+1-i|t-i}). \tag{7.17}$$

This approximation seems plausible, since in the exact predictor (7.16), $\theta_{t,i} \to \theta_i$. By iterating backwards, we can show that

$$\widehat{X}_{t+1|t} = \underbrace{\sum_{j=1}^{t-\max(p,q)} a_jX_{t+1-j}}_{\text{first part of AR}(\infty)\text{ expansion}} + \sum_{j=1}^{\max(p,q)} b_jX_j \tag{7.18}$$

where $|\gamma_j| \leq C\rho^t$, with $1/(1+\delta) < \rho < 1$ and the roots of $\theta(z)$ are outside $(1+\delta)$. On the other hand, the infinite predictor is

$$X_t(1) \;=\; \sum_{j=1}^{\infty} a_j X_{t+1-j} \quad (\text{since } X_{t+1} = \sum_{j=1}^{\infty} a_j X_{t+1-j} + \varepsilon_{t+1}).$$

Remark 7.8.1 *We prove (7.18) for the MA(1) model $X_t = \theta\varepsilon_{t-1} + \varepsilon_t$.The estimated predictor is*

$$\begin{aligned}
\widehat{X}_{t|t-1} &= \theta\left(X_{t-1} - \widehat{X}_{t-1|t-2}\right) \\
\Rightarrow X_t - \widehat{X}_{t|t-1} &= -\theta\left(X_{t-1} - \widehat{X}_{t-1|t-2}\right) + X_t \\
&= \sum_{j=0}^{t-1}(-\theta)^j X_{t-j-1} + (-\theta)^t\left(X_1 - \widehat{X}_{1|0}\right).
\end{aligned}$$

On the other hand, the infinite predictor is

$$\begin{aligned}
\widehat{X}_{t|t-1} &= \theta\left(X_{t-1} - \widehat{X}_{t-1|t-2}\right) \\
\Rightarrow X_t - \widehat{X}_{t|t-1} &= -\theta\left(X_{t-1} - \widehat{X}_{t-1|t-2}\right) + X_t \\
&= \sum_{j=0}^{t-1}(-\theta)^j X_{t-j-1} + (-\theta)^t\left(X_1 - \widehat{X}_{1|0}\right).
\end{aligned}$$

In summary, we have three one-step ahead predictors. The finite past best linear predictor:

$$X_{t+1|t} \;=\; \sum_{j=1}^{p}\phi_j X_{t+1-j} + \sum_{i=1}^{q}\theta_{i,t}(X_{t+1-i} - \widehat{X}_{t+1-i|t-i}) = \sum_{s=1}^{t}\phi_{t,s}X_{t+1-s} \tag{7.19}$$

The infinite past predictor:

$$X_t(1) \;=\; \sum_{j=1}^{p}\phi_j X_{t+1-j} + \sum_{i=1}^{q}\theta_i(X_{t+1-i} - X_{t-i}(1)) = \sum_{s=1}^{\infty}a_j X_{t+1-s} \tag{7.20}$$

and the approximate finite predictor:

$$\widehat{X}_{t+1|t} \;=\; \sum_{j=1}^{p}\phi_j X_{t+1-j} + \sum_{i=1}^{q}\theta_i(X_{t+1-i} - \widehat{X}_{t-i}(1)) = \sum_{s=1}^{t}a_j X_{t+1-s} + \sum_{s=1}^{\max(p,q)} b_s X_s. \tag{7.21}$$

These predictors will be very useful in deriving the approximate Gaussian likelihood for the ARMA model, see Section 9.2.2. We give a bound for the differences below.

Proposition 7.8.1 *Suppose $\{X_t\}$ is an ARMA process where the roots of $\phi(z)$ and $\theta(z)$ have roots which are greater in absolute value than $1+\delta$. Let $X_{t+1|t}$, $X_t(1)$ and $\widehat{X}_{t+1|t}$ be defined as in (7.19), (7.20) and (7.21) respectively. Then*

$$\mathrm{E}[\widehat{X}_{t+1|t} - X_t(1)]^2 \leq K\rho^t, \tag{7.22}$$

$$\mathrm{E}[X_{t+1|t} - X_t(1)]^2 \leq K\rho^t \tag{7.23}$$

and

$$\left|\mathrm{E}[X_{t+1} - X_{t+1|t}]^2 - \sigma^2\right| \leq K\rho^t \tag{7.24}$$

for any $\frac{1}{1+\delta} < \rho < 1$ and $\mathrm{var}(\varepsilon_t) = \sigma^2$.

7.9 ARMA models and the Kalman filter

7.9.1 The Kalman filter

The Kalman filter can be used to define a variant of the estimated predictor $\widehat{X}_t(1)$ described in (7.21). The Kalman filter construction is based on the state space equation

$$X_t = FX_{t-1} + V_t$$

where $\{X_t\}_t$ is an unobserved time series, F is a known matrix, $\mathrm{var}[V_t] = Q$ and $\{V_t\}_t$ are independent random variables that are independent of X_{t-1}. The observed equation

$$Y_t = HX_{t-1} + W_t$$

where $\{Y_t\}_t$ is the observed time series, $\mathrm{var}[W_t] = R$, $\{W_t\}_t$ are independent that are independent of X_{t-1}. Moreover $\{V_t\}_t$ and $\{W_t\}$ are jointly independent. The parameters can be made time-dependent, but this make the derivations notationally more cumbersome.

The standard notation is to let $\widehat{X}_{t+1|t} = P_{Y_1,\ldots,Y_t}(X_{t+1})$ and $P_{t+1|t} = \mathrm{var}[X_{t+1} - \widehat{X}_{t+1|t}]$ (predictive) and $\widehat{X}_{t+1|t+1} = P_{Y_1,\ldots,Y_t}(X_{t+1})$ and $P_{t+1|t+1} = \mathrm{var}[X_{t+1} - \widehat{X}_{t+1|t+1}]$ (update). The Kalman filter is an elegant method that iterates between the prediction steps $\widehat{X}_{t+1|t}$ and $P_{t+1|t}$ and the update steps $\widehat{X}_{t+1|t+1}$ and $P_{t+1|t+1}$. A proof is given at the end of the chapter. We summarise the algorithm below:

The Kalman equations

(i) <u>Prediction step</u> The conditional expectation

$$\widehat{X}_{t+1|t} = F\widehat{X}_{t|t}$$

and the corresponding mean squared error

$$P_{t+1|t} = FP_{t|t}F^{*} + Q.$$

(ii) <u>Update step</u> The conditional expectation

$$\widehat{X}_{t+1|t+1} = \widehat{X}_{t+1|t} + K_{t+1}\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right).$$

(note the appearance of Y_t, this is where the observed data plays a role in the prediction) where

$$K_{t+1} = P_{t+1|t}H^{*}[HP_{t+1|t}H^{*} + R]^{-1}$$

and the corresponding mean squared error

$$P_{t+1|t+1} = P_{t+1|t} - K_{t+1}HP_{t+1|t} = (I - K_{t+1}H)P_{t+1|t}.$$

(iii) There is also a smoothing step (which we ignore for now).

Thus we observe that if we can write a model in the above notation, then the predictors can be recursively updated. It is worth mentioning that in order to initiate the algorithm the initial values $X_{0|0}$ and $P_{0|0}$ are required.

7.9.2 The state space (Markov) representation of the ARMA model

There is no unique state-space representation of the ARMA model. We give below the elegant construction proposed in Akaike (1977) and expanded on in Jones (1980). This construction can be used as in prediction (via the Kalman filter) and to estimate the parameters in likelihood likelihood (but keep in mind initial conditions do matter). The construction is based on the best linear predictor of the infinite past.

We will assume $\{X_t\}$ has a causal ARMA(p,q) representation where

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t.$$

We now obtain a Markov-type representation of the above. It is based on best linear predictors given the infinite past. Let

$$X(t+r|t) = P_{X_t,X_{t-1},\ldots}(X_{t+r}),$$

where we recall that previously we used the notation $X_t(r) = X(t+r|t)$. The reason we change notation is to keep track of the time stamps. To obtain the representation we use that the ARMA model has the MA(∞) representation

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

where $\psi_0 = 1$. The MA(∞) coefficients can be derived from the ARMA parameters using the recursion

$$\psi_j = \theta_j + \sum_{k=1}^{j-1} \phi_k \theta_{j-k} \text{ for } j \geq 1,$$

setting the initial value $\psi_0 = 1$. Since $X(t+r|t)$ is the best linear predictor given the infinite past by using the results from Section 7.4 we have

$$\begin{aligned} X(t+r|t) &= P_{X_t,X_{t-1},\ldots}(X_{t+r}) = \sum_{j=r}^{\infty} \psi_{j+r} \varepsilon_{t+r-j} \\ X(t+r|t+1) &= P_{X_{t+1},X_t,X_{t-1},\ldots}(X_{t+r}) = \sum_{j=r-1}^{\infty} \psi_j \varepsilon_{t+r-j}. \end{aligned}$$

Thus taking differences we have

$$X(t+r|t+1) - X(t+r|t) = \psi_{r-1}\varepsilon_{t+1}.$$

Rewriting the above gives

$$X(t+r|t+1) = X(t+r|t) + \psi_{r-1}\varepsilon_{t+1}. \tag{7.25}$$

The simplest example of the above is $X_{t+r} = X(t+r|t+r) = X(t+r|t+r-1) + \varepsilon_{t+r}$. Based on (7.25) we have

$$\left(\begin{array}{c} X(t+1|t+1) \\ X(t+2|t+1) \\ X(t+3|t) \\ \vdots \\ X(t+r-1|t+1) \\ X(t+r|t+1) \end{array}\right) = \left(\begin{array}{cccccc} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ ? & ? & ? & \ldots & ? & ? \end{array}\right) \left(\begin{array}{c} X(t|t) \\ X(t+1|t) \\ X(t+2|t) \\ \vdots \\ X(t+r-2|t) \\ X(t+r-1|t) \end{array}\right) + \varepsilon_{t+1} \left(\begin{array}{c} 1 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{r-2} \\ \psi_{r-1} \end{array}\right).$$

The important observation is that the two vectors on the RHS of the above are independent, which is getting us towards a state space representation.

How to choose r in this representation and what are the ?s. Studying the last line in the above vector equation we note that

$$X(t+r|t+1) = X(t+r|t) + \psi_{r-1}\varepsilon_{t+1},$$

however $X(t+r|t)$ is not explicitly in the vector. Instead we need to find a linear combination of $X(t|t), \ldots, X(t+r-1|t)$ which gives $X(t+r|t)$. To do this we return to the ARMA representation

$$X_{t+r} = \sum_{j=1}^{p} \phi_j X_{t+r-j} + \sum_{i=1}^{q} \theta_i \varepsilon_{t+r-i} + \varepsilon_{t+r}.$$

The next part gets a little messy (you may want to look at Akaike or Jones for a better explanation).

Suppose that $r > q$, specifically let $r = q + 1$, then

$$\begin{aligned} P_{X_t,X_{t-1},\ldots}(X_{t+r}) &= \sum_{j=1}^{p} \phi_j P_{X_t,X_{t-1},\ldots}(X_{t+r-j}) + \underbrace{\sum_{i=1}^{q} \theta_i P_{X_t,X_{t-1},\ldots}(\varepsilon_{t+r-i}) + P_{X_t,X_{t-1},\ldots}(\varepsilon_{t+r})}_{\text{since } r>q \text{ this is } =0} \\ &= \sum_{j=1}^{p} \phi_j P_{X_t,X_{t-1},\ldots}(X_{t+r-j}). \end{aligned}$$

If, $p < q + 1$, then the above reduces to

$$X(t+r|t) = \sum_{j=1}^{p} \phi_j X(t+r-j|t).$$

If, on the other hand $p > r$, then

$$X(t+r|t) = \sum_{j=1}^{r} \phi_j X(t+r-j|t) + \sum_{j=r+1} \phi_j X_{t-j}.$$

Building $\{X_{t-j}\}_{j=1}^{r}$ from $\{X(t|t), \ldots, X(t+r-1|t)\}$ seems unlikely (it can probably proved it is not possible, but a proof escapes me for now). Thus, we choose $r \geq \max(p, q+1)$ (which will then gives everything in terms of the predictors). This choice gives

$$P_{X_t,X_{t-1},\ldots}(X_{t+r}) = \sum_{j=1}^{p} \phi_j X(t+r-j|t).$$

This allows us to construct the recursion equations for any $r \geq \max(p, q+1)$ by using the above to build the last row of the matrix. For simplicity we set $r = m = \max(p, q+1)$. If $p < \max(p, q+1)$, then for $p + 1 \leq r \leq m$ set $\phi_j = 0$. Define the recursion

$$\begin{pmatrix} X(t+1|t+1) \\ X(t+2|t+1) \\ X(t+3|t) \\ \vdots \\ X(t+m-1|t+1) \\ X(t+m|t+1) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ \phi_m & \phi_{m-1} & \phi_{m-2} & \ldots & \phi_2 & \phi_1 \end{pmatrix} \begin{pmatrix} X(t|t) \\ X(t+1|t) \\ X(t+2|t) \\ \vdots \\ X(t+m-2|t) \\ X(t+m-1|t) \end{pmatrix} + \varepsilon_{t+1} \begin{pmatrix} 1 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{m-2} \\ \psi_{m-1} \end{pmatrix}.$$

Let $\underline{Z}_t = (X(t|t), \ldots, X(t+m-1|t))$, and observe that $\underline{Z}_t$ is independent of ε_{t+1}. This yields the state space equation

$$\underline{Z}_{t+1} = F\underline{Z}_t + \underline{V}_{t+1}$$

where Φ is the matrix defined above and $\underline{V}'_{t+1} = \varepsilon_{t+1}(1, \psi_1, \ldots, \psi_{m-1}) = \varepsilon_{t+1}\underline{\psi}'_m$. By forward iterating

$$\underline{Z}_{t+1} = F\underline{Z}_t + \underline{V}_{t+1} \quad t \in \mathbb{Z}$$

from $t = -\infty$ the top entry of $\underline{Z}_t$ gives a stationary solution of the ARMA model. Of course in practice, we cannot start at $t = -\infty$ and start at $t = 0$, thus the initial conditions will play a role (and the solution won't precisely follow a stationary ARMA).

The observation model is

$$Y_{t+1} = (1, 0, \ldots, 0)\underline{Z}_{t+1},$$

where we note that $Y_{t+1} = X_{t+1}$. Thus we set $Y_t = X_t$ (where X_t is the observed time series).

7.9.3 Prediction using the Kalman filter

We use the Kalman filter described above where we set $Q = \mathrm{var}(\varepsilon_t)\underline{\psi}'_m\underline{\psi}_m$, $R = 0$, $H = (1, 0, \ldots, 0)$. This gives **The Kalman equations**

(1) Start with an initial value $\underline{Z}_{0|0}$. This part is where the approximation comes into play sinceY_0 is not observed. Typically a vectors of zeros are imputted for $\underline{Z}_{0|0}$ and recommendations for $P_{0|0}$ are given in given in Jones (1980) and Akaiki (1978). Then for $t > 0$ iterate on steps (2) and (3) below.

(2) Prediction step

$$\widehat{Z}_{t+1|t} = F\widehat{Z}_{t|t}$$

and the corresponding mean squared error

$$P_{t+1|t} = FP_{t|t}F^* + Q.$$

(3) Update step The conditional expectation

$$\widehat{Z}_{t+1|t+1} \quad = \quad \widehat{Z}_{t+1|t} + K_{t+1}\left(Y_{t+1} - H\widehat{Z}_{t+1|t}\right).$$

where

$$K_{t+1} = \frac{P_{t+1|t}H^*}{HP_{t+1|t}H^*}$$

and the corresponding mean squared error

$$P_{t+1|t+1} \quad = \quad P_{t+1|t} - K_tHP_{t+1|t} = (I - K_tH)P_{t+1|t}.$$

$\widehat{Z}_{t+1|t}$ will contain the linear predictors of $X_{t+1},\ldots,X_{t+m}$ given $X_1,\ldots,X_t$. They are "almost" the best linear predictors, but as in Section 7.8 the initial value plays a role (which is why it is only approximately the best linear predictor). Since we do not observe the infinite past we do not know $\underline{Z}_{m|m}$ (which is set to zero). The only way this can be exactly the best linear predictor is if $\underline{Z}_{m|m}$ were known, which it is not. Thus the approximate one-step ahead predictor is

$$X_{t+1|t} \approx [\underline{Z}_{t+1|t}]_{(1)} \approx \sum_{j=1}^{t} a_j X_{t-j},$$

where $\{a_j\}_{j=1}^{\infty}$ are the coefficients of the AR(∞) expansion corresponding to the ARMA model. The approximate r-step ahead predictor is $[\underline{Z}_{t+1|t}]_{(1)}$ (if $r \leq m$).

7.10 Forecasting for nonlinear models (advanced)

In this section we consider forecasting for nonlinear models. The forecasts we construct, may not necessarily/formally be the best linear predictor, because the best linear predictor is based on minimising the mean squared error, which we recall from Chapter 13 requires the existence of the higher order moments. Instead our forecast will be the conditional expection of X_{t+1} given the past (note that we can think of it as the best linear predictor). Furthermore, with the exception of the ARCH model we will derive approximation of the conditional expectation/best linear predictor, analogous to the forecasting approximation for the ARMA model, $\widehat{X}_{t+1|t}$ (given in (7.17)).

7.10.1 Forecasting volatility using an ARCH(p) model

We recall the ARCH(p) model defined in Section 13.2

$$X_t = \sigma_t Z_t \qquad \sigma_t^2 = a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2.$$

Using a similar calculation to those given in Section 13.2.1, we see that

$$\begin{aligned}
\mathrm{E}[X_{t+1}|X_t, X_{t-1}, \ldots, X_{t-p+1}] &= \mathrm{E}(Z_{t+1}\sigma_{t+1}|X_t, X_{t-1}, \ldots, X_{t-p+1}) = \underbrace{\sigma_{t+1}\mathrm{E}(Z_{t+1}|X_t, X_{t-1}, \ldots, X_{t-p+1})}_{\sigma_{t+1} \text{ function of } X_t,\ldots,X_{t-p+1}} \\
&= \sigma_{t+1} \underbrace{\mathrm{E}(Z_{t+1})}_{\text{by causality}} = 0 \cdot \sigma_{t+1} = 0.
\end{aligned}$$

In other words, past values of X_t have no influence on the expected value of X_{t+1}. On the other hand, in Section 13.2.1 we showed that

$$\mathrm{E}(X_{t+1}^2|X_t, X_{t-1}, \ldots, X_{t-p+1}) = \mathrm{E}(Z_{t+1}^2\sigma_{t+1}^2|X_t, X_{t-2}, \ldots, X_{t-p+1}) = \sigma_{t+1}^2\mathrm{E}[Z_{t+1}^2] = \sigma_{t+1}^2 = \sum_{j=1}^{p} a_j X_{t+1-j}^2,$$

thus X_t has an influence on the conditional mean squared/variance. Therefore, if we let $X_{t+k|t}$ denote the conditional variance of X_{t+k} given $X_t, \ldots, X_{t-p+1}$, it can be derived using the following recursion

$$\begin{aligned}
X_{t+1|t}^2 &= \sum_{j=1}^{p} a_j X_{t+1-j}^2 \\
X_{t+k|t}^2 &= \sum_{j=k}^{p} a_j X_{t+k-j}^2 + \sum_{j=1}^{k-1} a_j X_{t+k-j|k}^2 \quad 2 \leq k \leq p \\
X_{t+k|t}^2 &= \sum_{j=1}^{p} a_j X_{t+k-j|t}^2 \qquad k > p.
\end{aligned}$$

7.10.2 Forecasting volatility using a GARCH$(1, 1)$ model

We recall the GARCH$(1, 1)$ model defined in Section 13.3

$$\sigma_t^2 = a_0 + a_1 X_{t-1}^2 + b_1\sigma_{t-1}^2 = \left(a_1 Z_{t-1}^2 + b_1\right)\sigma_{t-1}^2 + a_0.$$

Similar to the ARCH model it is straightforward to show that $\mathrm{E}[X_{t+1}|X_t, X_{t-1}, \ldots] = 0$ (where we use the notation $X_t, X_{t-1}, \ldots$ to denote the infinite past or more precisely conditioned on the sigma algebra $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots)$). Therefore, like the ARCH process, our aim is to predict X_t^2.

We recall from Example 13.3.1 that if the GARCH the process is invertible (satisfied if $b < 1$), then

$$\mathrm{E}[X_{t+1}^2|X_t, X_{t-1}, \ldots] = \sigma_{t+1}^2 = a_0 + a_1 X_{t-1}^2 + b_1 \sigma_{t-1}^2 = \frac{a_0}{1-b} + a_1 \sum_{j=0}^{\infty} b^j X_{t-j}^2. \tag{7.26}$$

Of course, in reality we only observe the finite past $X_t, X_{t-1}, \ldots, X_1$. We can approximate $\mathrm{E}[X_{t+1}^2|X_t, X_{t-1}, \ldots, X_1]$ using the following recursion, set $\widehat{\sigma}_{1|0}^2 = 0$, then for $t \geq 1$ let

$$\widehat{\sigma}_{t+1|t}^2 = a_0 + a_1 X_t^2 + b_1 \widehat{\sigma}_{t|t-1}^2$$

(noting that this is similar in spirit to the recursive approximate one-step ahead predictor defined in (7.18)). It is straightforward to show that

$$\widehat{\sigma}_{t+1|t}^2 = \frac{a_0(1-b^{t+1})}{1-b} + a_1 \sum_{j=0}^{t-1} b^j X_{t-j}^2,$$

taking note that this is not the same as $\mathrm{E}[X_{t+1}^2|X_t, \ldots, X_1]$ (if the mean square error existed $\mathrm{E}[X_{t+1}^2|X_t, \ldots, X_1]$ would give a smaller mean square error), but just like the ARMA process it will closely approximate it. Furthermore, from (7.26) it can be seen that $\widehat{\sigma}_{t+1|t}^2$ closely approximates σ_{t+1}^2

Exercise 7.3 *To answer this question you need* R `install.package("tseries")` *then remember* `library("garch")`.

(i) You will find the Nasdaq data from 4th January 2010 - 15th October 2014 on my website.

(ii) By taking log differences fit a GARCH(1,1) model to the daily closing data (ignore the adjusted closing value) from 4th January 2010 - 30th September 2014 (use the function `garch(x, order = c(1, 1))` *fit the GARCH$(1,1)$ model).*

(iii) Using the fitted GARCH$(1,1)$ model, forecast the volatility σ_t^2 from October 1st-15th (noting that no trading is done during the weekends). Denote these forecasts as $\sigma_{t|0}^2$. Evaluate $\sum_{t=1}^{11} \sigma_{t|0}^2$

(iv) Compare this to the actual volatility $\sum_{t=1}^{11} X_t^2$ (where X_t are the log differences).

7.10.3 Forecasting using a BL$(1,0,1,1)$ model

We recall the Bilinear$(1,0,1,1)$ model defined in Section 13.4

$$X_t \quad = \quad \phi_1 X_{t-1} + b_{1,1} X_{t-1}\varepsilon_{t-1} + \varepsilon_t.$$

Assuming invertibility, so that ε_t can be written in terms of X_t (see Remark 13.4.2):

$$\varepsilon_t = \sum_{j=0}^{\infty} \left((-b)^j \prod_{i=0}^{j-1} X_{t-1-j} \right) [X_{t-j} - \phi X_{t-j-1}],$$

it can be shown that

$$X_t(1) = \mathrm{E}[X_{t+1}|X_t, X_{t-1}, \ldots] = \phi_1 X_t + b_{1,1} X_t \varepsilon_t.$$

However, just as in the ARMA and GARCH case we can obtain an approximation, by setting $\widehat{X}_{1|0} = 0$ and for $t \geq 1$ defining the recursion

$$\widehat{X}_{t+1|t} = \phi_1 X_t + b_{1,1} X_t \left(X_t - \widehat{X}_{t|t-1} \right).$$

See **?** and **?** for further details.

Remark 7.10.1 (How well does $\widehat{X}_{t+1|t}$ approximate $X_t(1)$?) *We now derive conditions for $\widehat{X}_{t+1|t}$ to be a close approximation of $X_t(1)$ when t is large. We use a similar technique to that used in Remark 7.8.1.*

We note that $X_{t+1} - X_t(1) = \varepsilon_{t+1}$ (since a future innovation, ε_{t+1}, cannot be predicted). We will show that $X_{t+1} - \widehat{X}_{t+1|t}$ is 'close' to ε_{t+1}. Subtracting $\widehat{X}_{t+1|t}$ from X_{t+1} gives the recursion

$$X_{t+1} - \widehat{X}_{t+1|t} = -b_{1,1}(X_t - \widehat{X}_{t|t-1})X_t + (b\varepsilon_t X_t + \varepsilon_{t+1}). \qquad (7.27)$$

We will compare the above recursion to the recursion based on ε_{t+1}. Rearranging the bilinear *equation gives*

$$\varepsilon_{t+1} = -b\varepsilon_t X_t + \underbrace{(X_{t+1} - \phi_1 X_t)}_{=b\varepsilon_t X_t + \varepsilon_{t+1}}. \qquad (7.28)$$

We observe that (7.27) and (7.28) are almost the same difference equation, the only difference is that an initial value is set for $\widehat{X}_{1|0}$. *This gives the difference between the two equations as*

$$\varepsilon_{t+1} - [X_{t+1} - \widehat{X}_{t+1|t}] = (-1)^t b^t X_1 \prod_{j=1}^{t} \varepsilon_j + (-1)^t b^t [X_1 - \widehat{X}_{1|0}] \prod_{j=1}^{t} \varepsilon_j.$$

Thus if $b^t \prod_{j=1}^t \varepsilon_j \overset{a.s.}{\to} 0$ *as* $t \to \infty$, *then* $\widehat{X}_{t+1|t} \overset{\mathcal{P}}{\to} X_t(1)$ *as* $t \to \infty$. *We now show that if* $\mathrm{E}[\log|\varepsilon_t| < -\log|b|$, *then* $b^t \prod_{j=1}^t \varepsilon_j \overset{a.s.}{\to} 0$. *Since* $b^t \prod_{j=1}^t \varepsilon_j$ *is a product, it seems appropriate to take logarithms to transform it into a sum. To ensure that it is positive, we take absolutes and t-roots*

$$\log|b^t \prod_{j=1}^{t} \varepsilon_j|^{1/t} = \log|b| + \underbrace{\frac{1}{t}\sum_{j=1}^{t} \log|\varepsilon_j|}_{\text{average of iid random variables}}.$$

Therefore by using the law of large numbers we have

$$\log|b^t \prod_{j=1}^{t} \varepsilon_j|^{1/t} = \log|b| + \frac{1}{t}\sum_{j=1}^{t} \log|\varepsilon_j| \overset{\mathcal{P}}{\to} \log|b| + \mathrm{E}\log|\varepsilon_0| = \gamma.$$

Thus we see that $|b^t \prod_{j=1}^t \varepsilon_j|^{1/t} \overset{a.s.}{\to} \exp(\gamma)$. *In other words,* $|b^t \prod_{j=1}^t \varepsilon_j| \approx \exp(t\gamma)$, *which will only converge to zero if* $\mathrm{E}[\log|\varepsilon_t| < -\log|b|$.

7.11 Nonparametric prediction (advanced)

In this section we briefly consider how prediction can be achieved in the nonparametric world. Let us assume that $\{X_t\}$ is a stationary time series. Our objective is to predict X_{t+1} given the past. However, we don't want to make any assumptions about the nature of $\{X_t\}$. Instead we want to obtain a predictor of X_{t+1} given X_t which minimises the means squared error, $\mathrm{E}[X_{t+1} - g(X_t)]^2$. It is well known that this is conditional expectation $\mathrm{E}[X_{t+1}|X_t]$. (since $\mathrm{E}[X_{t+1} - g(X_t)]^2 = \mathrm{E}[X_{t+1} -$

$\mathrm{E}(X_{t+1}|X_t)]^2 + \mathrm{E}[g(X_t) - \mathrm{E}(X_{t+1}|X_t)]^2)$. Therefore, one can estimate

$$\mathrm{E}[X_{t+1}|X_t = x] = m(x)$$

nonparametrically. A classical estimator of $m(x)$ is the Nadaraya-Watson estimator

$$\widehat{m}_n(x) = \frac{\sum_{t=1}^{n-1} X_{t+1} K(\frac{x-X_t}{b})}{\sum_{t=1}^{n-1} K(\frac{x-X_t}{b})},$$

where $K : \mathbb{R} \to \mathbb{R}$ is a kernel function (see Fan and Yao (2003), Chapter 5 and 6). Under some 'regularity conditions' it can be shown that $\widehat{m}_n(x)$ is a consistent estimator of $m(x)$ and converges to $m(x)$ in mean square (with the typical mean squared rate $O(b^4 + (bn)^{-1})$). The advantage of going the non-parametric route is that we have not imposed any form of structure on the process (such as linear/(G)ARCH/Bilinear). Therefore, we do not run the risk of misspecifying the model A disadvantage is that nonparametric estimators tend to be a lot worse than parametric estimators (in Chapter **??** we show that parametric estimators have $O(n^{-1/2})$ convergence which is faster than the nonparametric rate $O(b^2 + (bn)^{-1/2})$). Another possible disavantage is that if we wanted to include more past values in the predictor, ie. $m(x_1, \ldots, x_d) = \mathrm{E}[X_{t+1}|X_t = x_1, \ldots, X_{t-p} = x_d]$ then the estimator will have an extremely poor rate of convergence (due to the curse of dimensionality).

A possible solution to the problem is to assume some structure on the nonparametric model, and define a semi-parametric time series model. We state some examples below:

(i) An additive structure of the type

$$X_t = \sum_{j=1}^{p} g_j(X_{t-j}) + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid random variables.

(ii) A functional autoregressive type structure

$$X_t = \sum_{j=1}^{p} g_j(X_{t-d})X_{t-j} + \varepsilon_t.$$

(iii) The semi-parametric GARCH$(1,1)$

$$X_t = \sigma_t Z_t, \qquad \sigma_t^2 = b\sigma_{t-1}^2 + m(X_{t-1}).$$

However, once a structure has been imposed, conditions need to be derived in order that the model has a stationary solution (just as we did with the fully-parametric models).

See **?**, **?**, **?**, **?**, **?** etc.

7.12 The Wold Decomposition (advanced)

Section 5.5 nicely leads to the Wold decomposition, which we now state and prove. The Wold decomposition theorem, states that any stationary process, has something that appears close to an MA(∞) representation (though it is not). We state the theorem below and use some of the notation introduced in Section 5.5.

Theorem 7.12.1 *Suppose that $\{X_t\}$ is a second order stationary time series with a finite variance (we shall assume that it has mean zero, though this is not necessary). Then X_t can be uniquely expressed as*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t, \tag{7.29}$$

where $\{Z_t\}$ are uncorrelated random variables, with $\mathrm{var}(Z_t) = \mathrm{E}(X_t - X_{t-1}(1))^2$ (noting that $X_{t-1}(1)$ is the best linear predictor of X_t given $X_{t-1}, X_{t-2}, \ldots$) and $V_t \in \mathcal{X}_{-\infty} = \cap_{n=-\infty}^{-\infty} \mathcal{X}_n^{-\infty}$, where $\overline{\mathcal{X}_n^{-\infty}}$ is defined in (5.34).

PROOF. First let is consider the one-step ahead prediction of X_t given the infinite past, denoted $X_{t-1}(1)$. Since $\{X_t\}$ is a second order stationary process it is clear that $X_{t-1}(1) = \sum_{j=1}^{\infty} b_j X_{t-j}$, where the coefficients $\{b_j\}$ do not vary with t. For this reason $\{X_{t-1}(1)\}$ and $\{X_t - X_{t-1}(1)\}$ are second order stationary random variables. Furthermore, since $\{X_t - X_{t-1}(1)\}$ is uncorrelated with X_s for any $s \leq t$, then $\{X_s - X_{s-1}(1); s \in \mathbb{R}\}$ are uncorrelated random variables. Define $Z_s = X_s - X_{s-1}(1)$, and observe that Z_s is the one-step ahead prediction error. We recall from Section 5.5 that $X_t \in \overline{\mathrm{sp}}((X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots) \oplus \bar{sp}(\mathcal{X}_{-\infty}) = \oplus_{j=0}^{\infty} \overline{\mathrm{sp}}(Z_{t-j}) \oplus \bar{sp}(\mathcal{X}_{-\infty})$. Since the spaces $\oplus_{j=0}^{\infty} \overline{\mathrm{sp}}(Z_{t-j})$ and $\overline{\mathrm{sp}}(\mathcal{X}_{-\infty})$ are orthogonal, we shall first project X_t onto $\oplus_{j=0}^{\infty} \overline{\mathrm{sp}}(Z_{t-j})$, due to orthogonality the difference between X_t and its projection will be in $\overline{\mathrm{sp}}(\mathcal{X}_{-\infty})$. This will lead to the Wold decomposition.

First we consider the projection of X_t onto the space $\oplus_{j=0}^{\infty}\overline{\text{sp}}(Z_{t-j})$, which is

$$P_{Z_t,Z_{t-1},\ldots}(X_t) = \sum_{j=0}^{\infty}\psi_j Z_{t-j},$$

where due to orthogonality $\psi_j = cov(X_t,(X_{t-j} - X_{t-j-1}(1)))/\text{var}(X_{t-j} - X_{t-j-1}(1))$. Since $X_t \in \oplus_{j=0}^{\infty}\overline{\text{sp}}(Z_{t-j}) \oplus \bar{sp}(\mathcal{X}_{-\infty})$, the difference $X_t - P_{Z_t,Z_{t-1},\ldots}X_t$ is orthogonal to $\{Z_t\}$ and belongs in $\bar{sp}(\mathcal{X}_{-\infty})$. Hence we have

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

where $V_t = X_t - \sum_{j=0}^{\infty}\psi_j Z_{t-j}$ and is uncorrelated to $\{Z_t\}$. Hence we have shown (7.29). To show that the representation is unique we note that $Z_t, Z_{t-1},\ldots$ are an orthogonal basis of $\overline{\text{sp}}(Z_t, Z_{t-1},\ldots)$, which pretty much leads to uniqueness. □

Exercise 7.4 *Consider the process $X_t = A\cos(Bt + U)$ where A, B and U are random variables such that A, B and U are independent and U is uniformly distributed on $(0, 2\pi)$.*

(i) Show that X_t is second order stationary (actually it's stationary) and obtain its means and covariance function.

(ii) Show that the distribution of A and B can be chosen in such a way that $\{X_t\}$ has the same covariance function as the MA(1) process $Y_t = \varepsilon_t + \phi\varepsilon_t$ (where $|\phi| < 1$) (quite amazing).

(iii) Suppose A and B have the same distribution found in (ii).

(a) What is the best predictor of X_{t+1} given $X_t, X_{t-1},\ldots$?

(b) What is the best linear predictor of X_{t+1} given $X_t, X_{t-1},\ldots$?

It is worth noting that variants on the proof can be found in Brockwell and Davis (1998), Section 5.7 and Fuller (1995), page 94.

Remark 7.12.1 *Notice that the representation in (7.29) looks like an MA(∞) process. There is, however, a significant difference. The random variables $\{Z_t\}$ of an MA(∞) process are iid random variables and not just uncorrelated.*

We recall that we have already come across the Wold decomposition of some time series. In Section 6.4 we showed that a non-causal linear time series could be represented as a causal 'linear *time series' with uncorrelated but dependent innovations. Another example is in Chapter 13, where we explored ARCH/GARCH process which have an AR and ARMA type representation. Using this representation we can represent ARCH and GARCH processes as the weighted sum of* $\{(Z_t^2-1)\sigma_t^2\}$ *which are uncorrelated random variables.*

Remark 7.12.2 (Variation on the Wold decomposition) *In many technical proofs involving time series, we often use results related to the Wold decomposition. More precisely, we often decompose the time series in terms of an infinite sum of martingale differences. In particular, we define the sigma-algebra* $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots)$*, and suppose that* $\mathrm{E}(X_t|\mathcal{F}_{-\infty}) = \mu$*. Then by telescoping we can formally write* X_t *as*

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

where $Z_{t,j} = \mathrm{E}(X_t|\mathcal{F}_{t-j}) - \mathrm{E}(X_t|\mathcal{F}_{t-j-1})$*. It is straightforward to see that* $Z_{t,j}$ *are martingale differences, and under certain conditions (mixing, physical dependence, your favourite dependence flavour etc) it can be shown that* $\sum_{j=0}^{\infty}\|Z_{t,j}\|_p < \infty$ *(where* $\|\cdot\|_p$ *is the pth moment). This means the above representation holds almost surely. Thus in several proofs we can replace* $X_t - \mu$ *by* $\sum_{j=0}^{\infty} Z_{t,j}$*. This decomposition allows us to use martingale theorems to prove results.*

7.13 Kolmogorov's formula (advanced)

Suppose $\{X_t\}$ is a second order stationary time series. Kolmogorov's(-Szegö) theorem is an expression for the error in the linear prediction of X_t given the infinite past $X_{t-1}, X_{t-2}, \ldots$. It basically states that

$$\mathrm{E}\left[X_n - X_n(1)\right]^2 = \exp\left(\frac{1}{2\pi}\int_0^{2\pi} \log f(\omega) d\omega\right),$$

where f is the spectral density of the time series. Clearly from the definition we require that the spectral density function is bounded away from zero.

To prove this result we use (5.25);

$$\mathrm{var}[Y - \widehat{Y}] = \frac{\det(\Sigma)}{\det(\Sigma_{XX})}.$$

and Szegö's theorem (see, Gray's technical report, where the proof is given), which we state later on. Let $P_{X_1,\ldots,X_n}(X_{n+1}) = \sum_{j=1}^{n}\phi_{j,n}X_{n+1-j}$ (best linear predictor of X_{n+1} given $X_n,\ldots,X_1$). Then we observe that since $\{X_t\}$ is a second order stationary time series and using (5.25) we have

$$\mathrm{E}\left[X_{n+1} - \sum_{j=1}^{n}\phi_{n,j}X_{n+1-j}\right]^2 = \frac{\det(\Sigma_{n+1})}{\det(\Sigma_n)},$$

where $\Sigma_n = \{c(i-j); i,j = 0,\ldots,n-1\}$, and Σ_n is a non-singular matrix.

Szegö's theorem is a general theorem concerning Toeplitz matrices. Define the sequence of Toeplitz matrices $\Gamma_n = \{c(i-j); i,j = 0,\ldots,n-1\}$ and assume the Fourier transform

$$f(\omega) = \sum_{j\in\mathbb{Z}} c(j)\exp(ij\omega)$$

exists and is well defined ($\sum_j |c(j)|^2 < \infty$). Let $\{\gamma_{j,n}\}$ denote the Eigenvalues corresponding to Γ_n. Then for any function G we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}G(\gamma_{j,n}) \to \int_0^{2\pi}G(f(\omega))d\omega.$$

To use this result we return to $\mathrm{E}[X_{n+1} - \sum_{j=1}^{n}\phi_{n,j}X_{n+1-j}]^2$ and take logarithms

$$\begin{aligned}\log\mathrm{E}[X_{n+1} - \sum_{j=1}^{n}\phi_{n,j}X_{n+1-j}]^2 &= \log\det(\Sigma_{n+1}) - \log\det(\Sigma_n)\\ &= \sum_{j=1}^{n+1}\log\gamma_{j,n+1} - \sum_{j=1}^{n}\log\gamma_{j,n}\end{aligned}$$

where the above is because $\det\Sigma_n = \prod_{j=1}^{n}\gamma_{j,n}$ (where $\gamma_{j,n}$ are the eigenvalues of Σ_n). Now we apply Szegö's theorem using $G(x) = \log(x)$, this states that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\log(\gamma_{j,n}) \to \int_0^{2\pi}\log(f(\omega))d\omega.$$

thus for large n

$$\frac{1}{n+1}\sum_{j=1}^{n+1}\log\gamma_{j,n+1} \approx \frac{1}{n}\sum_{j=1}^{n}\log\gamma_{j,n}.$$

This implies that

$$\sum_{j=1}^{n+1} \log \gamma_{j,n+1} \approx \frac{n+1}{n} \sum_{j=1}^{n} \log \gamma_{j,n},$$

hence

$$\begin{aligned}
\log \mathrm{E}[X_{n+1} - \sum_{j=1}^{n} \phi_{n,j} X_{n+1-j}]^2 &= \log \det(\Sigma_{n+1}) - \log \det(\Sigma_n) \\
&= \sum_{j=1}^{n+1} \log \gamma_{j,n+1} - \sum_{j=1}^{n} \log \gamma_{j,n} \\
&\approx \frac{n+1}{n} \sum_{j=1}^{n} \log \gamma_{j,n} - \sum_{j=1}^{n} \log \gamma_{j,n} = \frac{1}{n} \sum_{j=1}^{n} \log \gamma_{j,n}.
\end{aligned}$$

Thus

$$\begin{aligned}
\lim_{n\to\infty} \log \mathrm{E}[X_{t+1} - \sum_{j=1}^{n} \phi_{n,j} X_{t+1-j}]^2 &= \lim_{n\to\infty} \log \mathrm{E}[X_{n+1} - \sum_{j=1}^{n} \phi_{n,j} X_{n+1-j}]^2 \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \log \gamma_{j,n} = \int_0^{2\pi} \log(f(\omega)) d\omega
\end{aligned}$$

and

$$\lim_{n\to\infty} \mathrm{E}[X_{t+1} - \sum_{j=1}^{n} \phi_{n,j} X_{t+1-j}]^2 = \exp\left(\int_0^{2\pi} \log(f(\omega)) d\omega \right).$$

This gives a rough outline of the proof. The precise proof can be found in Gray's technical report. There exists alternative proofs (given by Kolmogorov), see Brockwell and Davis (1998), Chapter 5.

This is the reason that in many papers the assumption

$$\int_0^{2\pi} \log f(\omega) d\omega > -\infty$$

is made. This assumption essentially ensures $X_t \notin \mathcal{X}_{-\infty}$.

Example 7.13.1 *Consider the AR(p) process $X_t = \phi X_{t-1} + \varepsilon_t$ (assume wlog that $|\phi| < 1$) where $\mathrm{E}[\varepsilon_t] = 0$ and $\mathrm{var}[\varepsilon_t] = \sigma^2$. We know that $X_t(1) = \phi X_t$ and*

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

We now show that

$$\exp\left(\frac{1}{2\pi}\int_0^{2\pi}\log f(\omega)d\omega\right) = \sigma^2. \tag{7.30}$$

We recall that the spectral density of the $AR(1)$ is

$$\begin{aligned} f(\omega) &= \frac{\sigma^2}{|1-\phi e^{i\omega}|^2} \\ \Rightarrow \log f(\omega) &= \log\sigma^2 - \log|1-\phi e^{i\omega}|^2. \end{aligned}$$

Thus

$$\frac{1}{2\pi}\int_0^{2\pi}\log f(\omega)d\omega = \underbrace{\frac{1}{2\pi}\int_0^{2\pi}\log\sigma^2 d\omega}_{=\log\sigma^2} - \underbrace{\frac{1}{2\pi}\int_0^{2\pi}\log|1-\phi e^{i\omega}|^2 d\omega}_{=0}.$$

There are various ways to prove that the second term is zero. Probably the simplest is to use basic results in complex analysis. By making a change of variables $z = e^{i\omega}$ we have

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi}\log|1-\phi e^{i\omega}|^2 d\omega &= \frac{1}{2\pi}\int_0^{2\pi}\log(1-\phi e^{i\omega})d\omega + \frac{1}{2\pi}\int_0^{2\pi}\log(1-\phi e^{-i\omega})d\omega \\ &= \frac{1}{2\pi}\int_0^{2\pi}\sum_{j=1}^{\infty}\left[\frac{\phi^j e^{ij\omega}}{j} + \frac{\phi^j e^{-ij\omega}}{j}\right]d\omega = 0. \end{aligned}$$

From this we immediately prove (7.30).

7.14 Appendix: Prediction coefficients for an AR(p) model

Define the p-dimension random vector $\underline{X}_t' = (X_t, \ldots, X_{t-p+1})$. We define the causal VAR(1) model in the vector form as

$$\underline{X}_t = \Phi\underline{X}_{t-1} + \underline{\varepsilon}_t$$

where $\underline{\varepsilon}_t' = (\varepsilon_t, 0, \ldots, 0)$ and

$$\Phi = \begin{pmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}. \tag{7.31}$$

Lemma 7.14.1 *Let Φ be defined as in (7.31) where parameters $\underline{\phi}$ are such that the roots of $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j$ lie outside the unit circle. Then*

$$[\Phi^{|\tau|+1} \underline{X}_p]_{(1)} = \sum_{\ell=1}^{p} X_\ell \sum_{s=0}^{p-\ell} \phi_{\ell+s} \psi_{|\tau|-s}. \tag{7.32}$$

where $\{\psi_j\}$ are the coefficients in the expansion $(1 - \sum_{j=1}^{p} \phi_j e^{-ij\omega})^{-1} = \sum_{j=0}^{\infty} \psi_s e^{-is\omega}$.

PROOF. The proof is based on the observation that the jth row of Φ^m ($m \geq 1$) is the $(j-1)$th row of Φ^{m-1} (due to the structure of A). Let $(\phi_{1,m}, \ldots, \phi_{p,m})$ denote the first row of Φ^m. Using this notation we have

$$\begin{pmatrix} \phi_{1,m} & \phi_{2,m} & \ldots & \phi_{p,m} \\ \phi_{1,m-1} & \phi_{2,m-1} & \ldots & \phi_{p,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{1,m-p+1} & \phi_{2,m-p+1} & \ldots & \phi_{p,m-p+1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} \phi_{1,m-1} & \phi_{2,m-1} & \ldots & \phi_{p,m-1} \\ \phi_{1,m-2} & \phi_{2,m-2} & \ldots & \phi_{p,m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{1,m-p} & \phi_{2,m-p} & \ldots & \phi_{p,m-p} \end{pmatrix}.$$

From the above we observe that $\phi_{\ell,m}$ satisfies the system of equations

$$\begin{aligned} \phi_{\ell,m} &= \phi_\ell \phi_{1,m-1} + \phi_{\ell+1,m-1} \qquad 1 \leq \ell \leq p-1 \\ \phi_{p,m} &= \phi_p \phi_{1,m-1}. \end{aligned} \tag{7.33}$$

Our aim is to obtain an expression for $\phi_{\ell,m}$ in terms of $\{\phi_j\}_{j=1}^{p}$ and $\{\psi_j\}_{j=0}^{\infty}$ which we now define. Since the roots of $\phi(\cdot)$ lies outside the unit circle the function $(1 - \sum_{j=1}^{p} \phi_j z^j)^{-1}$ is well defined for $|z| \leq 1$ and has the power series expansion $(1 - \sum_{i=1}^{p} \phi_i z)^{-1} = \sum_{i=0}^{\infty} \psi_i z^i$ for $|z| \leq 1$. We use the well know result $[\Phi^m]_{1,1} = \phi_{1,m} = \psi_m$. Using this we obtain an expression for the coefficients

$\{\phi_{\ell,m}; 2 \leq \ell \leq p\}$ in terms of $\{\phi_i\}$ and $\{\psi_i\}$. Solving the system of equations in (7.33), starting with $\phi_{1,1} = \psi_1$ and recursively solving for $\phi_{p,m}, \ldots, \phi_{2,m}$ we have

$$\begin{aligned}
\phi_{p,r} &= \phi_p \psi_{r-1} && m-p \leq r \leq m \\
\phi_{\ell,r} &= \phi_\ell \phi_{1,r-1} + \phi_{\ell+1,r-1} && 1 \leq \ell \leq p-1, \quad m-p \leq r \leq m
\end{aligned}$$

This gives $\phi_{p,m} = \phi_p \psi_{m-1}$, for $\ell = p-1$

$$\begin{aligned}
\phi_{p-1,m} &= \phi_{p-1}\phi_{1,m-1} + \phi_{p,m-1} \\
&= \phi_{p-1}\psi_{m-1} + \psi_p \psi_{m-2}
\end{aligned}$$

$$\begin{aligned}
\phi_{p-2,m} &= \phi_{p-2}\phi_{1,m-1} + \phi_{p-1,m-1} \\
&= \phi_{p-2}\psi_{m-1} + \phi_{p-1}\psi_{m-2} + \psi_p \psi_{m-3}
\end{aligned}$$

up to

$$\begin{aligned}
\phi_{1,m} &= \phi_1 \phi_{1,m-1} + \phi_{2,m-1} \\
&= \sum_{s=0}^{p-1} \phi_{1+s}\psi_{m-1-s} = (\psi_m).
\end{aligned}$$

This gives the general expression

$$\phi_{p-r,m} = \sum_{s=0}^{r} \phi_{p-r+s}\psi_{m-1-s} \qquad 0 \leq r \leq p-1.$$

In the last line of the above we change variables with $\ell = p - r$ to give for $m \geq 1$

$$\phi_{\ell,m} = \sum_{s=0}^{p-\ell} \phi_{\ell+s}\psi_{m-1-s} \qquad 1 \leq \ell \leq p,$$

where we set $\psi_0 = 1$ and for $t < 0$, $\psi_t = 0$. Therefore

$$[\Phi^{|\tau|+1}\underline{X}_p]_{(1)} = \sum_{\ell=1}^{p} X_\ell \sum_{s=0}^{p-\ell} \phi_{\ell+s}\psi_{|\tau|-s}.$$

Thus we obtain the desired result. □

A proof of Durbin-Levinson algorithm based on symmetric Toeplitz matrices

We now give an alternative proof which is based on properties of the (symmetric) Toeplitz matrix. We use (7.15), which is a matrix equation where

$$\Sigma_t \begin{pmatrix} \phi_{t,1} \\ \vdots \\ \phi_{t,t} \end{pmatrix} = \underline{r}_t, \tag{7.34}$$

with

$$\Sigma_t = \begin{pmatrix} c(0) & c(1) & c(2) & \ldots & c(t-1) \\ c(1) & c(0) & c(1) & \ldots & c(t-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c(t-1) & c(t-2) & \vdots & \vdots & c(0) \end{pmatrix} \quad \text{and} \quad \underline{r}_t = \begin{pmatrix} c(1) \\ c(2) \\ \vdots \\ c(t) \end{pmatrix}.$$

The proof is based on embedding $\underline{r}_{t-1}$ and Σ_{t-1} into Σ_{t-1} and using that $\Sigma_{t-1}\underline{\phi}_{t-1} = \underline{r}_{t-1}$.

To do this, we define the $(t-1)\times(t-1)$ matrix E_{t-1} which basically swops round all the elements in a vector

$$E_{t-1} = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & \vdots & 0 & 0 & 0 \end{pmatrix},$$

(recall we came across this swopping matrix in Section 6.2). Using the above notation, we have the interesting block matrix structure

$$\begin{aligned} \Sigma_t &= \begin{pmatrix} \Sigma_{t-1} & E_{t-1}\underline{r}_{t-1} \\ \underline{r}'_{t-1}E_{t-1} & c(0) \end{pmatrix} \\ \text{and } \underline{r}_t &= (\underline{r}'_{t-1}, c(t))'. \end{aligned}$$

Returning to the matrix equations in (7.34) and substituting the above into (7.34) we have

$$\Sigma_t \underline{\phi}_t = \underline{r}_t, \quad \Rightarrow \quad \begin{pmatrix} \Sigma_{t-1} & E_{t-1}\underline{r}_{t-1} \\ \underline{r}'_{t-1}E_{t-1} & c(0) \end{pmatrix} \begin{pmatrix} \underline{\phi}_{t-1,t} \\ \phi_{t,t} \end{pmatrix} = \begin{pmatrix} \underline{r}_{t-1} \\ c(t) \end{pmatrix},$$

where $\underline{\phi}'_{t-1,t} = (\phi_{1,t}, \ldots, \phi_{t-1,t})$. This leads to the two equations

$$\Sigma_{t-1}\underline{\phi}_{t-1,t} + E_{t-1}\underline{r}_{t-1}\phi_{t,t} = \underline{r}_{t-1} \tag{7.35}$$

$$\underline{r}'_{t-1}E_{t-1}\underline{\phi}_{t-1,t} + c(0)\phi_{t,t} = c(t). \tag{7.36}$$

We first show that equation (7.35) corresponds to the second equation in the Levinson-Durbin algorithm. Multiplying (7.35) by Σ_{t-1}^{-1}, and rearranging the equation we have

$$\underline{\phi}_{t-1,t} \quad = \quad \underbrace{\Sigma_{t-1}^{-1}\underline{r}_{t-1}}_{=\underline{\phi}_{t-1}} - \underbrace{\Sigma_{t-1}^{-1}E_{t-1}\underline{r}_{t-1}}_{=E_{t-1}\underline{\phi}_{t-1}}\phi_{t,t}.$$

Thus we have

$$\underline{\phi}_{t-1,t} = \underline{\phi}_{t-1} - \phi_{t,t}E_{t-1}\underline{\phi}_{t-1}. \tag{7.37}$$

This proves the second equation in Step 2 of the Levinson-Durbin algorithm.

We now use (7.36) to obtain an expression for $\phi_{t,t}$, which is the first equation in Step 1. Substituting (7.37) into $\underline{\phi}_{t-1,t}$ of (7.36) gives

$$\underline{r}'_{t-1}E_{t-1}\left(\underline{\phi}_{t-1} - \phi_{t,t}E_{t-1}\underline{\phi}_{t-1}\right) + c(0)\phi_{t,t} = c(t). \tag{7.38}$$

Thus solving for $\phi_{t,t}$ we have

$$\phi_{t,t} = \frac{c(t) - \underline{c}'_{t-1}E_{t-1}\underline{\phi}_{t-1}}{c(0) - \underline{c}'_{t-1}\underline{\phi}'_{t-1}}. \tag{7.39}$$

Noting that $r(t) = c(0) - \underline{c}'_{t-1}\underline{\phi}'_{t-1}$. (7.39) is the first equation of Step 2 in the Levinson-Durbin equation.

Note from this proof we do not need that the (symmetric) Toeplitz matrix is positive semi-definite. See Pourahmadi (2001), Chapter 7.

Prediction for ARMA models

Proof of equation (7.16) For the proof, we define the variables $\{W_t\}$, where $W_t = X_t$ for $1 \leq t \leq p$ and for $t > \max(p,q)$ let $W_t = \varepsilon_t + \sum_{i=1}^{q}\theta_i\varepsilon_{t-i}$ (which is the MA(q) part of the process). Since $X_{p+1} = \sum_{j=1}^{p}\phi_j X_{t+1-j} + W_{p+1}$ and so forth it is clear that $\overline{\text{sp}}(X_1,\ldots,X_t) = \overline{\text{sp}}(W_1,\ldots,W_t)$ (i.e. they are linear combinations of each other). To prove the result we use the following steps:

$$\begin{aligned}
P_{X_t,\ldots,X_1}(X_{t+1}) &= \sum_{j=1}^{p}\phi_j \underbrace{P_{X_t,\ldots,X_1}(X_{t+1-j})}_{X_{t+1-j}} + \sum_{i=1}^{q}\theta_i P_{X_t,\ldots,X_1}(\varepsilon_{t+1-i}) \\
&= \sum_{j=1}^{p}\phi_j X_{t+1-j} + \sum_{i=1}^{q}\theta_i \underbrace{P_{X_t-X_{t|t-1},\ldots,X_2-X_{2|1},X_1}(\varepsilon_{t+1-i})}_{=P_{W_t-W_{t|t-1},\ldots,W_2-W_{2|1},W_1}(\varepsilon_{t+1-i})} \\
&= \sum_{j=1}^{p}\phi_j X_{t+1-j} + \sum_{i=1}^{q}\theta_i P_{W_t-W_{t|t-1},\ldots,W_2-W_{2|1},W_1}(\varepsilon_{t+1-i}) \\
&= \sum_{j=1}^{p}\phi_j X_{t+1-j} + \sum_{i=1}^{q}\theta_i \underbrace{P_{W_{t+1-i}-W_{t+1-i|t-i},\ldots,W_t-W_{t|t-1}}(\varepsilon_{t+1-i})}_{\text{since } \varepsilon_{t+1-i}\text{is independent of } W_{t+1-i-j};j\geq 1} \\
&= \sum_{j=1}^{p}\phi_j X_{t+1-j} + \sum_{i=1}^{q}\theta_i \sum_{s=0}^{i-1} \underbrace{P_{W_{t+1-i+s}-W_{t+1-i+s|t-i+s}}(\varepsilon_{t+1-i})}_{\text{since } W_{t+1-i+s}-W_{t+1-i+s|t-i+s} \text{ are uncorrelated}} \\
&= \sum_{j=1}^{p}\phi_j X_{t+1-j} + \sum_{i=1}^{q}\theta_{t,i} \underbrace{(W_{t+1-i}-W_{t+1-i|t-i})}_{=X_{t+1-i}-X_{t+1-i|t-i}} \\
&= \sum_{j=1}^{p}\phi_j X_{t+1-j} + \sum_{i=1}^{q}\theta_{t,i}(X_{t+1-i}-X_{t+1-i|t-i}),
\end{aligned}$$

this gives the desired result.

We prove (7.18) for the ARMA(1, 2) model We first note that $\overline{\text{sp}}(X_1, X_t, \ldots, X_t) = \overline{\text{sp}}(W_1, W_2, \ldots, W_t)$, where $W_1 = X_1$ and for $t \geq 2$ $W_t = \theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\varepsilon_t$. The corresponding approximating predictor is defined as $\widehat{W}_{2|1} = W_1$, $\widehat{W}_{3|2} = W_2$ and for $t > 3$

$$\widehat{W}_{t|t-1} = \theta_1[W_{t-1} - \widehat{W}_{t-1|t-2}] + \theta_2[W_{t-2} - \widehat{W}_{t-2|t-3}].$$

Note that by using (7.17), the above is equivalent to

$$\underbrace{\widehat{X}_{t+1|t} - \phi_1 X_t}_{\widehat{W}_{t+1|t}} = \theta_1 \underbrace{[X_t - \widehat{X}_{t|t-1}]}_{=(W_t - \widehat{W}_{t|t-1})} + \theta_2 \underbrace{[X_{t-1} - \widehat{X}_{t-1|t-2}]}_{=(W_{t-1} - \widehat{W}_{t-1|t-2})}.$$

By subtracting the above from W_{t+1} we have

$$W_{t+1} - \widehat{W}_{t+1|t} = -\theta_1(W_t - \widehat{W}_{t|t-1}) - \theta_2(W_{t-1} - \widehat{W}_{t-1|t-2}) + W_{t+1}. \qquad (7.40)$$

It is straightforward to rewrite $W_{t+1} - \widehat{W}_{t+1|t}$ as the matrix difference equation

$$\underbrace{\left(\begin{array}{c} W_{t+1} - \widehat{W}_{t+1|t} \\ W_t - \widehat{W}_{t|t-1} \end{array}\right)}_{=\widehat{\underline{\varepsilon}}_{t+1}} = -\underbrace{\left(\begin{array}{cc} \theta_1 & \theta_2 \\ -1 & 0 \end{array}\right)}_{=Q}\underbrace{\left(\begin{array}{c} W_t - \widehat{W}_{t|t-1} \\ W_{t-1} - \widehat{W}_{t-1|t-2} \end{array}\right)}_{=\widehat{\underline{\varepsilon}}_t} + \underbrace{\left(\begin{array}{c} W_{t+1} \\ 0 \end{array}\right)}_{\underline{W}_{t+1}}$$

We now show that ε_{t+1} and $W_{t+1} - \widehat{W}_{t+1|t}$ lead to the same difference equation except for some initial conditions, it is this that will give us the result. To do this we write ε_t as function of $\{W_t\}$ (the irreducible condition). We first note that ε_t can be written as the matrix difference equation

$$\underbrace{\left(\begin{array}{c} \varepsilon_{t+1} \\ \varepsilon_t \end{array}\right)}_{=\underline{\varepsilon}_{t+1}} = -\underbrace{\left(\begin{array}{cc} \theta_1 & \theta_2 \\ -1 & 0 \end{array}\right)}_{Q}\underbrace{\left(\begin{array}{c} \varepsilon_t \\ \varepsilon_{t-1} \end{array}\right)}_{\underline{\varepsilon}_t} + \underbrace{\left(\begin{array}{c} W_{t+1} \\ 0 \end{array}\right)}_{\underline{W}_{t+1}} \qquad (7.41)$$

Thus iterating backwards we can write

$$\varepsilon_{t+1} = \sum_{j=0}^{\infty}(-1)^j[Q^j]_{(1,1)}W_{t+1-j} = \sum_{j=0}^{\infty}\tilde{b}_jW_{t+1-j},$$

where $\tilde{b}_j = (-1)^j[Q^j]_{(1,1)}$ (noting that $\tilde{b}_0 = 1$) denotes the $(1,1)$th element of the matrix Q^j (note we did something similar in Section **??**). Furthermore the same iteration shows that

$$\begin{aligned} \varepsilon_{t+1} &= \sum_{j=0}^{t-3}(-1)^j[Q^j]_{(1,1)}W_{t+1-j} + (-1)^{t-2}[Q^{t-2}]_{(1,1)}\varepsilon_3 \\ &= \sum_{j=0}^{t-3}\tilde{b}_jW_{t+1-j} + (-1)^{t-2}[Q^{t-2}]_{(1,1)}\varepsilon_3. \qquad (7.42) \end{aligned}$$

Therefore, by comparison we see that

$$\varepsilon_{t+1} - \sum_{j=0}^{t-3}\tilde{b}_jW_{t+1-j} = (-1)^{t-2}[Q^{t-2}\underline{\varepsilon}_3]_1 = \sum_{j=t-2}^{\infty}\tilde{b}_jW_{t+1-j}.$$

We now return to the approximation prediction in (7.40). Comparing (7.41) and (7.41) we see that they are almost the same difference equations. The only difference is the point at which the algorithm starts. $\underline{\varepsilon}_t$ goes all the way back to the start of time. Whereas we have set initial values for $\widehat{W}_{2|1} = W_1$, $\widehat{W}_{3|2} = W_2$, thus $\widehat{\underline{\varepsilon}}_3' = (W_3 - W_2, W_2 - W_1)$.Therefore, by iterating both (7.41) and (7.41) backwards, focusing on the first element of the vector and using (7.42) we have

$$\varepsilon_{t+1} - \widehat{\varepsilon}_{t+1} = \underbrace{(-1)^{t-2}[Q^{t-2}\underline{\varepsilon}_3]_1}_{=\sum_{j=t-2}^{\infty}\tilde{b}_j W_{t+1-j}} + (-1)^{t-2}[Q^{t-2}\widehat{\underline{\varepsilon}}_3]_1$$

We recall that $\varepsilon_{t+1} = W_{t+1} + \sum_{j=1}^{\infty}\tilde{b}_j W_{t+1-j}$ and that $\widehat{\varepsilon}_{t+1} = W_{t+1} - \widehat{W}_{t+1|t}$. Substituting this into the above gives

$$\widehat{W}_{t+1|t} - \sum_{j=1}^{\infty}\tilde{b}_j W_{t+1-j} = \sum_{j=t-2}^{\infty}\tilde{b}_j W_{t+1-j} + (-1)^{t-2}[Q^{t-2}\widehat{\underline{\varepsilon}}_3]_1.$$

Replacing W_t with $X_t - \phi_1 X_{t-1}$ gives (7.18), where the b_j can be easily deduced from $\tilde{b}_j$ and ϕ_1.

We now state a few results which will be useful later.

Lemma 7.14.2 *Suppose $\{X_t\}$ is a stationary time series with spectral density $f(\omega)$. Let $\boldsymbol{X}_t = (X_1, \ldots, X_t)$ and $\Sigma_t = \mathrm{var}(\boldsymbol{X}_t)$.*

(i) If the spectral density function is bounded away from zero (there is some $\gamma > 0$ such that $\inf_\omega f(\omega) > 0$), then for all t, $\lambda_{min}(\Sigma_t) \geq \gamma$ (where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest absolute eigenvalues of the matrix).

(ii) Further, $\lambda_{max}(\Sigma_t^{-1}) \leq \gamma^{-1}$.

(Since for symmetric matrices the spectral norm and the largest eigenvalue are the same, then $\|\Sigma_t^{-1}\|_{spec} \leq \gamma^{-1}$).

(iii) Analogously, $\sup_\omega f(\omega) \leq M < \infty$, then $\lambda_{\max}(\Sigma_t) \leq M$ (hence $\|\Sigma_t\|_{spec} \leq M$).

PROOF. See Chapter 10. □

Remark 7.14.1 *Suppose $\{X_t\}$ is an ARMA process, where the roots $\phi(z)$ and and $\theta(z)$ have absolute value greater than $1 + \delta_1$ and less than δ_2, then the spectral density $f(\omega)$ is bounded by $\mathrm{var}(\varepsilon_t)\frac{(1-\frac{1}{\delta_2})^{2p}}{(1-(\frac{1}{1+\delta_1})^{2p}} \leq f(\omega) \leq \mathrm{var}(\varepsilon_t)\frac{(1-(\frac{1}{1+\delta_1})^{2p}}{(1-\frac{1}{\delta_2})^{2p}}$. Therefore, from Lemma 7.14.2 we have that $\lambda_{\max}(\Sigma_t)$ and $\lambda_{\max}(\Sigma_t^{-1})$ is bounded uniformly over t.*

7.15 Appendix: Proof of the Kalman filter

In this section we prove the recursive equations used to define the Kalman filter. The proof is straightforward and used the multi-stage projection described in Section 5.1.4 (which has been already been used to prove the Levinson-Durbin algorithm and forms the basis of the Burg algorithm).

The Kalman filter construction is based on the state space equation

$$X_t = FX_{t-1} + V_t$$

where $\{X_t\}_t$ is an unobserved time series, F is a known matrix, $\mathrm{var}[V_t] = Q$ and $\{V_t\}_t$ are independent random variables that are independent of X_{t-1}. The observed equation

$$Y_t = HX_{t-1} + W_t$$

where $\{Y_t\}_t$ is the observed time series, $\mathrm{var}[W_t] = R$, $\{W_t\}_t$ are independent that are independent of X_{t-1}. Moreover $\{V_t\}_t$ and $\{W_t\}$ are jointly independent. The parameters can be made time-dependent, but this make the derivations notationally more cumbersome.

The derivation of the Kalman equations are based on the projections discussed in Section 5.3. In particular, suppose that X, Y, Z are random variables then

$$P_{Y,Z}(X) \quad = \quad P_Y(X) + \alpha_X(Z - P_Y(Z)) \tag{7.43}$$

where

$$\alpha_X = \frac{\mathrm{cov}(X, Z - P_Y(Z))}{\mathrm{var}(Z - P_Y(Z))}$$

and

$$\mathrm{var}[X - P_{Y,Z}(X)] = \mathrm{cov}[X, X - P_{Y,Z}(X)], \tag{7.44}$$

these properties we have already used a number of time.

The standard notation is $\widehat{X}_{t+1|t} = P_{Y_1,\ldots,Y_t}(X_{t+1})$ and $P_{t+1|t} = \mathrm{var}[X_{t+1} - \widehat{X}_{t+1|t}]$ (predictive) and $\widehat{X}_{t+1|t+1} = P_{Y_1,\ldots,Y_t}(X_{t+1})$ and $P_{t+1|t+1} = \mathrm{var}[X_{t+1} - \widehat{X}_{t+1|t+1}]$ (update).

The Kalman equations

(i) Prediction step

The conditional expectation

$$\widehat{X}_{t+1|t} = F\widehat{X}_{t|t}$$

and the corresponding mean squared error

$$P_{t+1|t} = FP_{t|t}F^{*} + Q.$$

(ii) Update step

The conditional expectation

$$\widehat{X}_{t+1|t+1} \quad = \quad \widehat{X}_{t+1|t} + K_{t+1}\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right).$$

where

$$K_{t+1} = P_{t+1|t}H^{*}[HP_{t+1|t}H^{*} + R]^{-1}$$

and the corresponding mean squared error

$$P_{t+1|t+1} \quad = \quad P_{t+1|t} - K_t HP_{t+1|t} = (I - K_t H)P_{t+1|t}$$

(iii) There is also a smoothing step (which we ignore for now).

The Kalman filter iteratively evaluates step (i) and (ii) for $t = 2, 3, \ldots$. We start with $\widehat{X}_{t-1|t-1}$ and $P_{t-1|t-1}$.

Derivation of predictive equations The best linear predictor:

$$\begin{aligned}\widehat{X}_{t+1|t} = P_{Y_1,\ldots,Y_t}(X_{t+1}) \quad &= \quad P_{Y_1,\ldots,Y_t}(FX_t + V_{t+1}) \\ &= \quad P_{Y_1,\ldots,Y_t}(FX_t) + P_{Y_1,\ldots,Y_t}(V_{t+1}) = FP_{Y_1,\ldots,Y_t}(X_t) = F\widehat{X}_{t|t}.\end{aligned}$$

The mean squared error

$$\begin{aligned}
P_{t+1|t} &= \mathrm{var}[X_{t+1}-\widehat{X}_{t+1|t}] = \mathrm{var}[FX_t+V_{t+1}-F\widehat{X}_{t|t}] \\
&= \mathrm{var}[F(X_t-\widehat{X}_{t|t})+V_{t+1}] \\
&= \mathrm{var}[F(X_t-\widehat{X}_{t|t})]+\mathrm{var}[V_{t+1}] \\
&= F\mathrm{var}[X_t-\widehat{X}_{t|t}]F^*+\mathrm{var}[V_{t+1}] = FP_{t|t}F^*+Q.
\end{aligned}$$

This gives the two predictors from the previous update equations. Next the update equations (which is slightly more tricky).

<u>Derivation of the update equations</u> Now we expand the projection space from $\mathrm{sp}(Y_1,\ldots,Y_t)$ to $\mathrm{sp}(Y_1,\ldots,Y_t,Y_{t+1})$. But as the recursion uses $\mathrm{sp}(Y_1,\ldots,Y_t)$ we represent

$$\mathrm{sp}(Y_1,\ldots,Y_t,Y_{t+1}) = \mathrm{sp}(Y_1,\ldots,Y_t,Y_{t+1}-P_{Y_1,\ldots,Y_t}(Y_{t+1})).$$

Note that

$$\begin{aligned}
Y_{t+1}-P_{Y_1,\ldots,Y_t}(Y_{t+1}) &= Y_{t+1}-P_{Y_1,\ldots,Y_t}(HX_{t+1}+W_{t+1}) \\
&= Y_{t+1}-H\widehat{X}_{t+1|t}.
\end{aligned}$$

Thus by using (7.43) we have

$$\widehat{X}_{t+1|t+1} = P_{Y_1,\ldots,Y_t,Y_{t+1}}(X_{t+1}) = \widehat{X}_{t+1|t}+\alpha\left(Y_{t+1}-H\widehat{X}_{t+1|t}\right)$$

where

$$\alpha = \mathrm{var}(Y_{t+1}-H\widehat{X}_{t+1|t})^{-1}\mathrm{cov}(X_{t+1},Y_{t+1}-H\widehat{X}_{t+1|t}).$$

We now find an expression for $\alpha = K_{t+1}$ (K_{t+1} is the typical notation). We recall that $Y_{t+1} = HX_{t+1}+W_{t+1}$, thus $Y_{t+1}-H\widehat{X}_{t+1|t} = H(X_{t+1}-X_{t+1|t})+W_{t+1}$. Thus

$$\begin{aligned}
\mathrm{cov}(X_{t+1},Y_{t+1}-H\widehat{X}_{t+1|t}) &= \mathrm{cov}(X_{t+1},H(X_{t+1}-X_{t+1|t})+W_{t+1}) \\
&= \mathrm{cov}(X_{t+1},H(X_{t+1}-X_{t+1|t})) = \mathrm{cov}(X_{t+1}-X_{t+1|t},X_{t+1})H^* \\
&= \mathrm{var}(X_{t+1}-X_{t+1|t}) = P_{t+1|t}H^* \qquad (7.45)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}(Y_{t+1} - H\widehat{X}_{t+1|t}) &= \mathrm{var}(H(X_{t+1} - X_{t+1|t}) + W_{t+1}) \\
&= H\mathrm{var}(X_{t+1} - X_{t+1|t})H^{*} + \mathrm{var}(W_{t+1}) \\
&= HP_{t+1|t}H^{*} + R.
\end{aligned}$$

Therefore, altogether

$$K_{t+1} = P_{t+1|t}H^{*}[HP_{t+1|t}H^{*} + R]^{-1}$$

$$\widehat{X}_{t+1|t+1} = \widehat{X}_{t+1|t} + K_{t+1}\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right).$$

Often K_{t+1} or $K_{t+1}\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right)$ is referred to as the Kalman gain, which the "gain" when including the additional term Y_{t+1} in the prediction. Finally we calculate the variance. Again using (7.44) we have

$$\begin{aligned}
P_{t+1|t+1} &= \mathrm{var}[X_{t+1} - \widehat{X}_{t+1|t+1}] = \mathrm{cov}[X_{t+1}, X_{t+1} - \widehat{X}_{t+1|t+1}] \\
&= \mathrm{cov}\left[X_{t+1}, X_{t+1} - \widehat{X}_{t+1|t} - K_{t}\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right)\right] \\
&= \mathrm{cov}\left[X_{t+1}, X_{t+1} - \widehat{X}_{t+1|t}\right] - \mathrm{cov}\left[X_{t+1}, K_{t}\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right)\right] \\
&= P_{t+1|t} - K_{t}HP_{t+1|t} = (I - K_{t}H)P_{t+1|t}
\end{aligned}$$

where the above follows from (7.45). I have a feeling the above may be a little wrong in terms of of brackets.

Chapter 8

Estimation of the mean and covariance

Objectives

- To derive the sample autocovariance of a time series, and show that this is a positive definite sequence.
- To show that the variance of the sample covariance involves fourth order cumulants, which can be unwielding to estimate in practice. But under linearity the expression for the variance greatly simplifies.
- To show that under linearity the correlation does not involve the fourth order cumulant. This is the Bartlett formula.
- To use the above results to construct a test for uncorrelatedness of a time series (the Portmanteau test). And understand how this test may be useful for testing for independence in various different setting. Also understand situations where the test may fail.

Here we summarize the Central limit theorems we will use in this chapter. The simplest is the case of iid random variables. The first is the classical central limit theorem. Suppose that $\{X_i\}$ are iid random variables with mean μ and variance $\sigma^2 < \infty$. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \overset{\mathcal{D}}{\to} \mathcal{N}(0,\sigma^2).$$

A small variant on the classical CLT is the case that $\{X_i\}$ are independent random variables (but not identically distributed). Suppose $\mathrm{E}[X_i] = \mu_i$, $\mathrm{var}[X_i] = \sigma_i^2 < \infty$ and for every $\varepsilon > 0$

$$\frac{1}{s_n^2}\sum_{i=1}^{n} \mathrm{E}\left((X_i - \mu_i)^2 I(s_n^{-1}|X_i - \mu_i| > \varepsilon)\right) \to 0$$

where $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$, which is the variance of $\sum_{i=1}^{n} X_i$ (the above condition is called the Lindeberg condition). Then

$$\frac{1}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \sum_{i=1}^{n} (X_i - \mu_i) \overset{\mathcal{D}}{\to} \mathcal{N}(0,1).$$

The Lindeberg condition looks unwieldy, however by using Chebyshev's and Hölder inequality it can be reduced to simple bounds on the moments.

Remark 8.0.1 (The aims of the Lindeberg condition) *The Lindeberg condition essential requires a uniform bound in the tails for all the random variables $\{X_i\}$ in the sum. For example, suppose X_i are t-distributed random variables where X_i is distributed with a t-distribution with $(2+i^{-1})$ degrees of freedom. We know that the number of df (which can be non-integer-valued) gets thicker the lower the df. Furthermore, $\mathrm{E}[X_i^2] < \infty$ only if X_i has a df greater than 2. Therefore, the second moments of X_i exists. But as i gets larger, X_i has thicker tails. Making it impossible (I believe) to find a uniform bound such that Lindeberg's condition is satisified.*

Note that the Lindeberg condition generalizes to the conditional Lindeberg condition when dealing with martingale differences.

We now state a generalisation of this central limit to triangular arrays. Suppose that $\{X_{t,n}\}$ are independent random variables with mean zero. Let $S_n = \sum_{t=1}^{n} X_{t,n}$ we assume that $\mathrm{var}[S_n] = \sum_{t=1}^{n} \mathrm{var}[X_{t,n}] = 1$. For example, in the case that $\{X_t\}$ are iid random variables and $S_n = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}[X_t - \mu] = \sum_{t=1}^{n} X_{t,n}$, where $X_{t,n} = \sigma^{-1}n^{-1/2}(X_t - \mu)$. If for all $\varepsilon > 0$

$$\sum_{t=1}^{n} \mathrm{E}\left(X_{t,n}^2 I(|X_{t,n}| > \varepsilon)\right) \to 0,$$

then $S_n \overset{\mathcal{D}}{\to} \mathcal{N}(0,1)$.

8.1 An estimator of the mean

Suppose we observe $\{Y_t\}_{t=1}^n$, where

$$Y_t = \mu + X_t,$$

where μ is the finite mean, $\{X_t\}$ is a zero mean stationary time series with absolutely summable covariances ($\sum_k |\mathrm{cov}(X_0, X_k)| < \infty$). Our aim is to estimate the mean μ. The most obvious estimator is the sample mean, that is $\bar{Y}_n = n^{-1}\sum_{t=1}^n Y_t$ as an estimator of μ.

8.1.1 The sampling properties of the sample mean

We recall from Example 3.3.1 that we obtained an expression for the sample mean. We showed that

$$\mathrm{var}(\bar{Y}_n) = \frac{1}{n}c(0) + \frac{2}{n}\sum_{k=1}^{n}(\frac{n-k}{n})c(k).$$

Furthermore, if $\sum_k |c(k)| < \infty$, then in Example 3.3.1 we showed that

$$\mathrm{var}(\bar{Y}_n) \approx \frac{1}{n}c(0) + \frac{2}{n}\sum_{k=1}^{\infty}c(k).$$

Thus if the time series has sufficient decay in its correlation structure a mean squared consistent estimator of the sample mean can be achieved. However, one drawback is that the dependency means that one observation will influence the next, and if the influence is positive (seen by a positive covariance), the resulting estimator may have a (much) larger variance than the iid case.

The above result does not require any more conditions on the process, besides second order stationarity and summability of its covariance. However, to obtain confidence intervals we require a stronger result, namely a central limit theorem for the sample mean. The above conditions are not enough to give a central limit theorem. To obtain a CLT for sums of the form $\sum_{t=1}^n X_t$ we need the following main ingredients:

(i) The variance needs to be finite.

(ii) The dependence between X_t decreases the further apart in time the observations. However, this is more than just the correlation, it really means the dependence.

The above conditions are satisfied by linear time series, if the cofficients ϕ_j decay sufficient fast. However, these conditions can also be verified for nonlinear time series (for example the (G)ARCH and Bilinear model described in Chapter 13).

We now state the asymptotic normality result for linear models.

Theorem 8.1.1 *Suppose that X_t is a linear time series, of the form $X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}$, where ε_t are iid random variables with mean zero and variance one, $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $\sum_{j=-\infty}^{\infty} \psi_j \neq 0$. Let $Y_t = \mu + X_t$, then we have*

$$\sqrt{n}\left(\bar{Y}_n - \mu\right) = \mathcal{N}(0, V)$$

where $V = c(0) + 2\sum_{k=1}^{\infty} c(k)$.

PROOF. Later in this course we will give precise details on how to prove asymptotic normality of several different type of estimators in time series. However, we give a small flavour here by showing asymptotic normality of $\bar{Y}_n$ in the special case that $\{X_t\}_{t=1}^{n}$ satisfy an MA(q) model, then explain how it can be extended to MA(∞) processes.

The main idea of the proof is to transform/approximate the average into a quantity that we know is asymptotic normal. We know if $\{\epsilon_t\}_{t=1}^{n}$ are iid random variables with mean μ and variance one then

$$\sqrt{n}(\bar{\epsilon}_n - \mu) \overset{\mathcal{D}}{\to} \mathcal{N}(0, 1). \tag{8.1}$$

We aim to use this result to prove the theorem. Returning to $\bar{Y}_n$ by a change of variables ($s = t - j$) we can show that

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} Y_t &= \mu + \frac{1}{n}\sum_{t=1}^{n} X_t = \mu + \frac{1}{n}\sum_{t=1}^{n}\sum_{j=0}^{q} \psi_j \varepsilon_{t-j} \\
&= \mu + \frac{1}{n}\sum_{s=1}^{n-q} \varepsilon_s \left(\sum_{j=0}^{q} \psi_j\right) + \sum_{s=-q+1}^{0} \varepsilon_s \left(\sum_{j=q-s}^{q} \psi_j\right) + \sum_{s=n-q+1}^{n} \varepsilon_s \left(\sum_{j=0}^{n-s} \psi_j\right) \\
&= \mu + \frac{n-q}{n}\left(\sum_{j=0}^{q} \psi_j\right) \frac{1}{n-q}\sum_{s=1}^{n-q} \varepsilon_s + \frac{1}{n}\sum_{s=-q+1}^{0} \varepsilon_s \left(\sum_{j=q+s}^{q} \psi_j\right) + \frac{1}{n}\sum_{s=n-q+1}^{n} \varepsilon_s \left(\sum_{j=0}^{n-s} \psi_j\right) \\
&:= \mu + \frac{(n-q)\Psi}{n} \bar{\varepsilon}_{n-q} + E_1 + E_2,
\end{aligned} \tag{8.2}$$

where $\Psi = \sum_{j=0}^{q} \psi_j$. It is straightforward to show that $\mathrm{E}|E_1| \leq Cn^{-1}$ and $\mathrm{E}|E_2| \leq Cn^{-1}$.

Finally we examine $\frac{(n-q)\Psi}{n}\bar{\varepsilon}_{n-q}$. We note that if the assumptions are not satisfied and $\sum_{j=0}^{q} \psi_j = 0$ (for example the process $X_t = \varepsilon_t - \varepsilon_{t-1}$), then

$$\frac{1}{n}\sum_{t=1}^{n} Y_t \;=\; \mu + \frac{1}{n}\sum_{s=-q+1}^{0} \varepsilon_s \left(\sum_{j=q-s}^{q} \psi_j\right) + \frac{1}{n}\sum_{s=n-q+1}^{n} \varepsilon_s \left(\sum_{j=0}^{n-s} \psi_j\right).$$

This is a degenerate case, since E_1 and E_2 only consist of a finite number of terms and thus if ε_t are non-Gaussian these terms will never be asymptotically normal. Therefore, in this case we simply have that $\frac{1}{n}\sum_{t=1}^{n} Y_t = \mu + O(\frac{1}{n})$ (this is why in the assumptions it was stated that $\Psi \neq 0$).

On the other hand, if $\Psi \neq 0$, then the dominating term in $\bar{Y}_n$ is $\bar{\varepsilon}_{n-q}$. From (8.1) it is clear that $\sqrt{n-q}\bar{\varepsilon}_{n-q} \overset{\mathcal{P}}{\to} \mathcal{N}(0,1)$ as $n \to \infty$. However, for finite q, $\sqrt{(n-q)/n} \overset{\mathcal{P}}{\to} 1$, therefore $\sqrt{n}\bar{\varepsilon}_{n-q} \overset{\mathcal{P}}{\to} \mathcal{N}(0,1)$. Altogether, substituting $\mathrm{E}|E_1| \leq Cn^{-1}$ and $\mathrm{E}|E_2| \leq Cn^{-1}$into (8.2) gives

$$\sqrt{n}\left(\bar{Y}_n - \mu\right) = \Psi\sqrt{n}\bar{\varepsilon}_{n-q} + O_p(\frac{1}{n}) \overset{\mathcal{P}}{\to} \mathcal{N}\left(0, \Psi^2\right).$$

With a little work, it can be shown that $\Psi^2 = V$.

Observe that the proof simply approximated the sum by a sum of iid random variables. In the case that the process is a MA(∞) or linear time series, a similar method is used. More precisely, we have

$$\begin{aligned}\sqrt{n}\left(\bar{Y}_n - \mu\right) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j} = \frac{1}{\sqrt{n}}\sum_{j=0}^{\infty}\psi_j\sum_{s=1-j}^{n-j}\varepsilon_s \\ &= \frac{1}{\sqrt{n}}\sum_{j=0}^{\infty}\psi_j\sum_{t=1}^{n}\varepsilon_t + R_n\end{aligned}$$

where

$$\begin{aligned}R_n &= \frac{1}{\sqrt{n}}\sum_{j=0}^{\infty}\psi_j\left(\sum_{s=1-j}^{n-j}\varepsilon_s - \sum_{s=1}^{n}\varepsilon_s\right) \\ &= \frac{1}{\sqrt{n}}\sum_{j=0}^{n}\psi_j\left(\sum_{s=1-j}^{0}\varepsilon_s - \sum_{s=n-j}^{n}\varepsilon_s\right) + \frac{1}{\sqrt{n}}\sum_{j=n+1}^{\infty}\psi_j\left(\sum_{s=1-j}^{n-j}\varepsilon_s - \sum_{s=1}^{n}\varepsilon_s\right) \\ &:= R_{n1} + R_{n2} + R_{n3} + R_{n4}.\end{aligned}$$

We will show that $\mathrm{E}[R_{n,j}^2] = o(1)$ for $1 \leq j \leq 4$. We start with $R_{n,1}$

$$\begin{aligned}
\mathrm{E}[R_{n,1}^2] &= \frac{1}{n}\sum_{j_1,j_2=0}^{n}\psi_{j_1}\psi_{j_2}\mathrm{cov}\left(\sum_{s_1=1-j_1}^{0}\varepsilon_{s_1}, \sum_{s_2=1-j_2}^{0}\varepsilon_{s_2}\right) \\
&= \frac{1}{n}\sum_{j_1,j_2=0}^{n}\psi_{j_1}\psi_{j_2}\min[j_1-1, j_2-1] \\
&= \frac{1}{n}\sum_{j=0}^{n}\psi_j^2(j-1) + \frac{2}{n}\sum_{j_1=0}^{n}\psi_{j_1}, \sum_{j_2=0}^{j_1-1}\psi_{j_2}\min[j_2-1] \\
&\leq \frac{1}{n}\sum_{j=0}^{n}\psi_j^2(j-1) + \frac{2\Psi}{n}\sum_{j_1=0}^{n}|j_1\psi_{j_1}|.
\end{aligned}$$

Since $\sum_{j=0}^{\infty}|\psi_j| < \infty$ and, thus, $\sum_{j=0}^{\infty}|\psi_j|^2 < \infty$, then by dominated convegence $\sum_{j=0}^{n}[1-j/n]\psi_j \to \sum_{j=0}^{\infty}\psi_j$ and $\sum_{j=0}^{n}[1-j/n]\psi_j^2 \to \sum_{j=0}^{\infty}\psi_j^2$ as $n \to \infty$. This implies that $\sum_{j=0}^{n}(j/n)\psi_j \to 0$ and $\sum_{j=0}^{n}(j/n)\psi_j^2 \to 0$. Substituting this into the above bounds for $\mathrm{E}[R_{n,1}^2]$ we immediately obtain $\mathrm{E}[R_{n,1}^2] = o(1)$. Using the same argument we obtain the same bound for $R_{n,2}, R_{n,3}$ and $R_{n,4}$. Thus

$$\sqrt{n}\left(\bar{Y}_n - \mu\right) = \Psi\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_t + o_p(1)$$

and the result then immediately follows. □

Estimation of the so called long run variance (given in Theorem 8.1.1) can be difficult. There are various methods that can be used, such as estimating the spectral density function (which we define in Chapter 10) at zero. Another approach proposed in Lobato (2001) and Shao (2010) is to use the method of so called self-normalization which circumvents the need to estimate the long run mean, by privotalising the statistic.

8.2 An estimator of the covariance

Suppose we observe $\{Y_t\}_{t=1}^n$, to estimate the covariance we can estimate the covariance $c(k) = \mathrm{cov}(Y_0, Y_k)$ from the the observations. A plausible estimator is

$$\hat{c}_n(k) = \frac{1}{n}\sum_{t=1}^{n-|k|}(Y_t - \bar{Y}_n)(Y_{t+|k|} - \bar{Y}_n), \tag{8.3}$$

since $\mathrm{E}[(Y_t - \bar{Y}_n)(Y_{t+|k|} - \bar{Y}_n)] \approx c(k)$. Of course if the mean of Y_t is known to be zero ($Y_t = X_t$), then the covariance estimator is

$$\widehat{c}_n(k) = \frac{1}{n}\sum_{t=1}^{n-|k|} X_t X_{t+|k|}. \tag{8.4}$$

The eagle-eyed amongst you may wonder why we don't use $\frac{1}{n-|k|}\sum_{t=1}^{n-|k|} X_t X_{t+|k|}$, when $\hat{c}_n(k)$ is a biased estimator, whereas $\frac{1}{n-|k|}\sum_{t=1}^{n-|k|} X_t X_{t+|k|}$ is not. However $\hat{c}_n(k)$ has some very nice properties which we discuss in the lemma below. The sample autocorrelation is the ratio

$$\widehat{\rho}_n(r) = \frac{\widehat{c}_n(r)}{\widehat{c}_n(0)}.$$

Most statistical software will have functions that evaluate the sample autocorrelation.

Lemma 8.2.1 *Suppose we define the empirical covariances*

$$\widehat{c}_n(k) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-|k|} X_t X_{t+|k|} & |k| \leq n-1 \\ 0 & \textit{otherwise} \end{cases}$$

then $\{\widehat{c}_n(k)\}$ is a positive definite sequence. Therefore, using Lemma 3.4.1 there exists a stationary time series $\{Z_t\}$ which has the covariance $\hat{c}_n(k)$.

PROOF. There are various ways to show that $\{\hat{c}_n(k)\}$ is a positive definite sequence. One method uses that the spectral density corresponding to this sequence is non-negative, we give this proof in Section 10.4.1.

Here we give an alternative proof. We recall a sequence is semi-positive definite if for any vector $\underline{a} = (a_1, \ldots, a_r)'$ we have

$$\sum_{k_1,k_2=1}^{r} a_{k_1} a_{k_2} \hat{c}_n(k_1 - k_2) = \sum_{k_1,k_2=1}^{n} a_{k_1} a_{k_2} \hat{c}_n(k_1 - k_2) = \underline{a}'\widehat{\Sigma}_n \underline{a} \geq 0$$

where

$$\widehat{\Sigma}_n = \begin{pmatrix} \hat{c}_n(0) & \hat{c}_n(1) & \hat{c}_n(2) & \ldots & \hat{c}_n(n-1) \\ \hat{c}_n(1) & \hat{c}_n(0) & \hat{c}_n(1) & \ldots & \hat{c}_n(n-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{c}_n(n-1) & \hat{c}_n(n-2) & \vdots & \vdots & \hat{c}_n(0) \end{pmatrix},$$

noting that $\hat{c}_n(k) = \frac{1}{n}\sum_{t=1}^{n-|k|} X_t X_{t+|k|}$. However, $\hat{c}_n(k) = \frac{1}{n}\sum_{t=1}^{n-|k|} X_t X_{t+|k|}$ has a very interesting construction, it can be shown that the above convariance matrix is $\widehat{\Sigma}_n = \mathbf{X}_n\mathbf{X}_n'$, where $\mathbf{X}_n$ is a $n \times 2n$ matrix with

$$\mathbf{X}_n = \begin{pmatrix} 0 & 0 & \ldots & 0 & X_1 & X_2 & \ldots & X_{n-1} & X_n \\ 0 & 0 & \ldots & X_1 & X_2 & \ldots & X_{n-1} & X_n & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_1 & X_2 & \ldots & X_{n-1} & X_n & 0 & \ldots & \ldots & 0 \end{pmatrix}$$

Using the above we have

$$\underline{a}'\widehat{\Sigma}_n\underline{a} = \underline{a}'\mathbf{X}_n\mathbf{X}_n'\underline{a} = \|\mathbf{X}'\underline{a}\|_2^2 \geq 0.$$

This this proves that $\{\hat{c}_n(k)\}$ is a positive definite sequence.

Finally, by using Theorem 3.4.1, there exists a stochastic process with $\{\hat{c}_n(k)\}$ as its autocovariance function. □

8.2.1 Asymptotic properties of the covariance estimator

The main reason we construct an estimator is either for testing or constructing a confidence interval for the parameter of interest. To do this we need the variance and distribution of the estimator. It is impossible to derive the finite sample distribution, thus we look at their asymptotic distribution. Besides showing asymptotic normality, it is important to derive an expression for the variance.

In an ideal world the variance will be simple and will not involve unknown parameters. Usually in time series this will not be the case, and the variance will involve several (often an infinite) number of parameters which are not straightforward to estimate. Later in this section we show that the variance of the sample covariance can be extremely complicated. However, a substantial simplification can arise if we consider only the sample correlation (not variance) and assume linearity of the time series. This result is known as Bartlett's formula (you may have come across Maurice Bartlett before, besides his fundamental contributions in time series he is well known for proposing the famous Bartlett correction). This example demonstrates, how the assumption of linearity can really simplify problems in time series analysis and also how we can circumvent certain problems in which arise by making slight modifications of the estimator (such as going from covariance to correlation).

The following theorem gives the asymptotic sampling properties of the covariance estimator (8.3). One proof of the result can be found in Brockwell and Davis (1998), Chapter 8, Fuller (1995), but it goes back to Bartlett (indeed its called Bartlett's formula). We prove the result in Section **??**.

Theorem 8.2.1 *Suppose $\{X_t\}$ is a mean zero linear stationary time series where*

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

where $\sum_j |\psi_j| < \infty$, $\{\varepsilon_t\}$ are iid random variables with $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{E}(\varepsilon_t^4) < \infty$. Suppose we observe $\{X_t : t = 1, \ldots, n\}$ and use (8.3) as an estimator of the covariance $c(k) = \mathrm{cov}(X_0, X_k)$. Define $\hat{\rho}_n(r) = \hat{c}_n(r)/\hat{c}_n(0)$ as the sample correlation. Then for each $h \in \{1, \ldots, n\}$

$$\sqrt{n}(\hat{\rho}_n(h) - \rho(h)) \overset{\mathcal{D}}{\rightarrow} \mathcal{N}(0, W_h) \tag{8.5}$$

where $\hat{\rho}_n(h) = (\hat{\rho}_n(1), \ldots, \hat{\rho}_n(h))$, $\rho(h) = (\rho(1), \ldots, \rho(h))$ and

$$\begin{aligned}(W_h)_{ij} &= \sum_{k=-\infty}^{\infty} \Big\{ \rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) \\ &\quad -2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i) \Big\}. \end{aligned} \tag{8.6}$$

Equation (8.6) is known as Bartlett's formula.

In Section 8.3 we apply the method for checking for correlation in a time series. We first show how the expression for the asymptotic variance is obtained.

8.2.2 The asymptotic properties of the sample autocovariance and autocorrelation

In order to show asymptotic normality of the autocovariance and autocorrelation we require the following result. For any coefficients $\{\alpha_{r_j}\}_{j=0}^{d} \in \mathbb{R}^{d+1}$ (such that σ_α^2, defined below, is non-zero) we have

$$\sqrt{n}\left(\sum_{j=0}^{d} \alpha_{r_j} \frac{1}{n} \sum_{t=1}^{n-|r_j|} X_t X_{t+r_j} - \sum_{j=0}^{d} \alpha_{r_j} c(r_j) \right) \overset{\mathcal{D}}{\rightarrow} \mathcal{N}\left(0, \sigma_\alpha^2\right), \tag{8.7}$$

for some $\sigma_\alpha^2 < \infty$. This result can be proved under a whole host of conditions including

- The time series is linear, $X_t = \sum_j \psi_j \varepsilon_{t-j}$, where $\{\varepsilon_t\}$ are iid, $\sum_j |\psi_j| < \infty$ and $\mathrm{E}[\varepsilon_t^4] < \infty$.
- α and β-mixing with sufficient mixing rates and moment conditions (which are linked to the mixing rates).
- Physical dependence
- Other dependence measures.

All these criterions essentially show that the time series $\{X_t\}$ becomes "increasingly independent" the further apart the observations are in time. How this dependence is measured depends on the criterion, but it is essential for proving the CLT. We do not prove the above. Our focus in this section will be on the variance of the estimator.

Theorem 8.2.2 *Suppose that condition (8.7) is satisfied (and $\sum_{h\in\mathbb{Z}} |c(h)| < \infty$ and $\sum_{h_1,h_2,h_3} |\kappa_4(h_1,h_2,h_3)| < \infty$; this is a cumulant, which we define in the section below), then*

$$\sqrt{n}\begin{pmatrix} \widehat{c}_n(0) - c(0) \\ \widehat{c}_n(r_1) - c(r_1) \\ \vdots \\ \widehat{c}_n(r_d) - c(r_d) \end{pmatrix} \overset{\mathcal{P}}{\rightarrow} \mathcal{N}\left(0, V_{d+1}\right)$$

where

$$\begin{aligned} (V_{d+1})_{i,j} &= \sum_{k=-\infty}^{\infty} c(k)c(k+r_{i-1}-r_{j-1}) + \sum_{k=-\infty}^{\infty} c(k+r_{i-1}-1)c(k-r_{j-1}-1) + \\ &\quad \sum_{k=-\infty}^{\infty} \kappa_4(r_{i-1}-1, k, k+r_{j-1}-1) \end{aligned} \tag{8.8}$$

where we set $r_0 = 0$.

PROOF. The first part of the proof simply follows from (8.7). The derivation for V_{d+1} is given in Section 8.2.3, below. □

In order to prove the results below, we partition V_{d+1} into a term which contains the covariances and the term which contains the fourth order cumulants (which we have yet to define). Let $V_{d+1} =$

$C_{d+1} + K_{d+1}$, where

$$\begin{aligned}
(C_{d+1})_{i,j} &= \sum_{k=-\infty}^{\infty} c(k)c(k + r_{i-1} - r_{j-1}) + \sum_{k=-\infty}^{\infty} c(k + r_{i-1})c(k - r_{j-1}) \\
(K_{d+1})_{i,j} &= \sum_{k=-\infty}^{\infty} \kappa_4(r_{i-1}, k, k + r_{j-1}).
\end{aligned} \tag{8.9}$$

and set $r_0 = 0$. So far we have not defined κ_4. However, it is worth bearing in mind that if the time series $\{X_t\}$ is Gaussian, then this term is zero i.e. $K_{d+1} = 0$. Thus estimation of the variance of the sample covariance for Gaussian time series is relatively straightforward as it only depends on the covariance.

We now derive the sampling properties of the sample autocorrelation.

Lemma 8.2.2 *Suppose that conditions in Theorem 8.2.2 hold. Then*

$$\sqrt{n} \begin{pmatrix} \widehat{\rho}_n(r_1) - \rho(r_1) \\ \vdots \\ \widehat{\rho}_n(r_d) - \rho(r_d) \end{pmatrix} \stackrel{\mathcal{P}}{\rightarrow} \mathcal{N}\left(0, G(C_{d+1} + K_{d+1})G'\right)$$

where $r_j \neq 0$, C_{d+1} and K_{d+1} are defined as in equation (8.9) and G is a $d \times (d+1)$ dimensional matrix where

$$G = \frac{1}{c(0)} \begin{pmatrix} -\rho(r_1) & 1 & 0 & \ldots & 0 \\ -\rho(r_2) & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \ddots & 0 \\ -\rho(r_d) & 0 & \ldots & \ldots & 1 \end{pmatrix}$$

PROOF. We define the $g : \mathbb{R}^{d+1} \to \mathbb{R}^d$ vector function

$$g(x_0, x_1, \ldots, x_d) = \left(\frac{x_1}{x_0}, \ldots, \frac{x_d}{x_0} \right).$$

We observe that $(\widehat{\rho}(r_1), \ldots, \widehat{\rho}(r_d)) = g(\widehat{c}_n(0), \widehat{c}_n(r_1), \ldots, \widehat{c}_n(r_d))$. Thus

$$\nabla g(c(0), \ldots, c(r_d)) = \begin{pmatrix} -\frac{c(r_1)}{c(0)^2} & \frac{1}{c(0)} & 0 & \ldots & 0 \\ -\frac{c(r_2)}{c(0)^2} & 0 & \frac{1}{c(0)} & \ldots & 0 \\ \vdots & \vdots & \ldots & \ddots & 0 \\ -\frac{c(r_d)}{c(0)^2} & 0 & \ldots & \ldots & \frac{1}{c(0)} \end{pmatrix} = G.$$

Therefore, by using Theorem 8.2.2 together with the continuous mapping theorem we obtain the result. □

Comparing Theorem 8.2.2 to the asymptotically pivotal result $\sqrt{n}\underline{\rho}_{h,n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_h)$ in (**??**) it is clear that additional assumptions are required for the result to be pivotal. Therefore, in the following theorem we consider the case that $\{X_t\}$ is a linear time series, which includes the special case that $\{X_t\}$ are iid random variables. First, we make some observations about G and $GC_{d+1}G'$. Note that the assumption of linearity of a time series can be checked (see, for example, Subba Rao and Gabr (1980)).

Remark 8.2.1 *(i) Basic algebra gives*

$$\begin{aligned}(GC_{d+1}G')_{r_1,r_2} &= \sum_{k=-\infty}^{\infty}\Big\{\rho(k+r_1)\rho(k+r_2) + \rho(k-r_1)\rho(k+r_2) + 2\rho(r_1)\rho(r_2)\rho^2(k) \\ &\quad -2\rho(r_1)\rho(k)\rho(k+r_2) - 2\rho(r_2)\rho(k)\rho(k+r_1)\Big\}. \qquad (8.10)\end{aligned}$$

(ii) Though it may not seem directly relevant. It is easily seen that the null space of the matrix G is

$$\mathcal{N}(G) = \big\{\alpha \underline{c}_{d+1}; \alpha \in \mathbb{R}\big\}$$

where $\underline{c}'_{d+1} = (c(0), c(r_1), \ldots, c(r_d))$. This property will be useful in proving Bartlett's formula (below).

Theorem 8.2.3 *Suppose $\{X_t\}$ is a mean zero <u>linear</u> stationary time series where*

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

with $\sum_j |\psi_j| < \infty$, $\{\varepsilon_t\}$ are iid random variables with $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{E}(\varepsilon_t^4) < \infty$. Suppose we observe $\{X_t : t = 1, \ldots, n\}$ and use (8.3) as an estimator of the covariance $c(k) = \mathrm{cov}(X_0, X_k)$. Then we have

$$\sqrt{n} \begin{pmatrix} \widehat{\rho}_n(r_1) - \rho(r_1) \\ \vdots \\ \widehat{\rho}_n(r_d) - \rho(r_d) \end{pmatrix} \xrightarrow{\mathcal{P}} \mathcal{N}\left(0, GC_{d+1}G'\right),$$

where an explicit expression for $GC_{d+1}G'$ is given in (8.10) (this is called Bartlett's formula).

PROOF. To prove the result we use Lemma 8.2.2. However, we observe that the term $GK_{d+1}G'$ has disappeared. In Section 8.2.3 we show that for (univariate) linear processes $GK_{d+1}G' = 0$. □

Remark 8.2.2
- *Under linearity of the time series, Brockwell and Davis (2002), Theorem 7.2.2 show that the above theorem also holds for linear time series whose fourth moment does not exist. This result requires slightly stronger assumptions on the coefficients $\{\psi_j\}$.*
- *This allusive fourth cumulant term does not disappear for vector linear processes.*

Using Theorem 8.2.3, we can prove (**??**) for iid time series. Since iid random variables are a special case of a linear time series ($\phi_j = 0$ for all $j \neq 0$) with $c(r) = 0$ for all $r \neq 0$. Substituting this into Theorem 8.2.3 gives

$$\sqrt{n} \begin{pmatrix} \widehat{\rho}_n(1) \\ \vdots \\ \widehat{\rho}_n(h) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_h).$$

Using this result we obtain the critical values in the ACF plots and the Box-Pierce test. However, from Lemma 8.2.2 we observe that the results can be misleading for time series which are uncorrelated but not necessarily iid. Before discussing this, we first prove the above results. These calculations are a little tedious, but they are useful in understanding how to deal with many different types of statistics of a time series (not just the sample autocovariances).

8.2.3 The covariance of the sample autocovariance

Our aim in this section is to derive an expression for $\mathrm{cov}\left(\widehat{c}_n(r_1), \widehat{c}_n(r_2)\right)$. To simply notation we focus on the variance ($r_1 = r_2$), noting that the same calculations carry over to the covariance.

Approach 1 Use the moment expansion of a covariance

$$\begin{aligned}
\mathrm{var}[\widehat{c}_n(r)] &= \frac{1}{n^2}\sum_{t,\tau=1}^{n-|r|}\mathrm{cov}(X_tX_{t+r},X_\tau X_{\tau+r}) \\
&= \frac{1}{n^2}\sum_{t,\tau=1}^{n-|r|}(\mathrm{E}(X_tX_{t+r},X_\tau X_{\tau+r})-\mathrm{E}(X_tX_{t+r})\mathrm{E}(X_\tau X_{\tau+r})) \\
&= \frac{1}{n^2}\sum_{t,\tau=1}^{n-|r|}\left(\mathrm{E}(X_tX_{t+r},X_\tau X_{\tau+r})-c(r)^2\right).
\end{aligned}$$

Studying the above and comparing it to the expansion of $\mathrm{var}(\bar{X})$ when the $\{X_t\}$ are iid, we would expect that $\mathrm{var}[\widehat{c}_n(r)] = O(n^{-1})$. But it is difficult to see what is happening with this expansion. Though it is possible to use this method. We use an alternative expansion in terms of cumulants.

Approach 2 Use an expansion of the covariance of products in terms of products of cumulants. Suppose A, B, C and D are zero mean (real) random variables. Then

$$\underbrace{\mathrm{cov}}_{=\mathrm{cum}}(AB,CD) = \underbrace{\mathrm{cov}}_{=\mathrm{cum}}(A,C)\underbrace{\mathrm{cov}}_{=\mathrm{cum}}(B,D) + \underbrace{\mathrm{cov}}_{=\mathrm{cum}}(A,D)\underbrace{\mathrm{cov}}_{=\mathrm{cum}}(B,C) + \mathrm{cum}(A,B,C,D). \quad (8.11)$$

This result can be generalized to higher order cumulants, see Brillinger (2001).

Below, we formally define a cumulant and explain why it is a useful tool in time series.

Background: What are cumulants?

To understand what they are and why they are used, we focus the following discussion for fourth order cumulants.

The joint cumulant of $X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}$ (denoted as $\mathrm{cum}(X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3})$) is the coefficient of the term $s_1s_2s_3s_4$ in the power series expansion of

$$K(s_1,s_2,s_3,s_4) = \log \mathrm{E}[e^{is_1X_t+is_2X_{t+k_1}+is_3X_{t+k_2}+is_4X_{t+k_4}}].$$

Thus

$$\mathrm{cum}(X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}) = \frac{\partial^4 K(s_1,s_2,s_3,s_4)}{\partial s_1\partial s_1\partial s_3\partial s_4}\rfloor_{s_1,s_2,s_3,s_4=0}$$

It looks very similar to the definition of moments and there is a one to one correpondence between the moments and the cumulants. It can be shown that the cumulant corresponding to coefficient of $s_i s_j$ is $\mathrm{cum}(X_{t+k_i}, X_{t+k_j})$ (the covariance is often called the second order cumulant).

Properties

- If X_t is independent of $X_{t+k_1}, X_{t+k_2}, X_{t+k_3}$ then

  $$\mathrm{cum}\,(X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}) = 0.$$

  This is because the log of the corresponding characteristic function is

  $$\log \mathrm{E}[e^{is_1X_t+is_2X_{t+k_1}+is_3X_{t+k_2}+is_4X_{t+k_4}}] = \log \mathrm{E}[e^{is_1X_t}] + \log[\mathrm{E}[e^{is_2X_{t+k_1}+is_3X_{t+k_2}+is_4X_{t+k_4}}].$$

  Differentiating the above with respect to $s_1s_2s_3s_4$ gives zero.

- If $X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}$ is multivariate Gaussian, then all cumulants higher than order 2 are zero. This is easily seen, by recalling that the characteristic function of a multivariate normal distribution is

  $$C(s_1, s_2, s_3, s_4) = \exp(i\underline{\mu}'\underline{s} - \frac{1}{2}\underline{s}'\Sigma\underline{s})$$

  where $\underline{\mu}$ and Σ are the mean and variance of $X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}$ respectively. Based on the above, we observe that $\log C(s_1, s_2, s_3, s_4)$ is an order two multivariate polynomial.

  Note that this property can be used to prove CLTs.

- Cumulants satisfy the follow multilinear property

  $$\begin{aligned} & \mathrm{cum}\,(aX_1 + bY_1 + c, X_2, X_3, X_4) \\ = \;& a\mathrm{cum}\,(X_1, X_2, X_3, X_4) + b\mathrm{cum}\,(Y_1, X_2, X_3, X_4) \end{aligned}$$

  where a, b and c are scalars.

- The influence of stationarity:

  From the definition of the characteristic function, if the time series $\{X_t\}$ is strictly stationary. Then

  $$\log \mathrm{E}[e^{is_1X_t+is_2X_{t+k_1}+is_3X_{t+k_2}+is_4X_{t+k_4}}] = \log \mathrm{E}[e^{is_1X_0+is_2X_{k_1}+is_3X_{k_2}+is_4X_{k_4}}].$$

Thus, analogous to covariances, cumulants are invariant to shift

$$\mathrm{cum}(X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}) = \mathrm{cum}(X_0, X_{k_1}, X_{k_2}, X_{k_3}) = \kappa_4(k_1, k_2, k_3).$$

Comparisons between the covariance and higher order cumulants

(a) The covariance is invariant to ordering $\mathrm{cov}[X_t, X_{t+k}] = \mathrm{cov}[X_{t+k}, X_t]$.

Like the covariance, the joint cumulant $\mathrm{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}]$ is also invariant to order.

(b) The covariance $\mathrm{cov}[X_t, X_{t+k}]$ is a measure of linear dependence between X_t and X_{t+k}.

The cumulant is measuring the dependence between $\mathrm{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}]$ in "three directions" (though as far as I am aware, unlike the covariance it has no clear geometric interpretation). For example, if $\{X_t\}$ is a zero mean time series then

$$\begin{aligned} & \mathrm{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}] \\ = \; & \mathrm{E}[X_t X_{t+k_1} X_{t+k_2} X_{t+k_3}] - \mathrm{E}[X_t X_{t+k_1}]\mathrm{E}[X_{t+k_2} X_{t+k_3}] \\ & -\mathrm{E}[X_t X_{t+k_2}]\mathrm{E}[X_{t+k_1} X_{t+k_3}] - \mathrm{E}[X_t X_{t+k_3}]\mathrm{E}[X_{t+k_1} X_{t+k_2}]. \end{aligned} \tag{8.12}$$

Unlike the covariance, the cumulants do not seem to satisfy any non-negative definite conditions.

(c) In time series we usually assume that the covariance decays over time i.e. if $k > 0$

$$|\mathrm{cov}[X_t, X_{t+k}]| \leq \alpha(k)$$

where $\alpha(k)$ is a positive sequence such that $\sum_k \alpha(k) < \infty$. This can easily be proved for linear time series with $\sum_j |\psi_j| < \infty$[1].

For a large class of time series, the analogous result is true for cumulants. I.e. if $k_1 \leq k_2 \leq k_3$

[1] This is easily shown by noting that if $X_t = \sum_j \psi_j \varepsilon_{t-j}$ then $\mathrm{cov}(X_t, X_{t+h}) = \sigma^2 \sum_j \psi_j \psi_{j+h}$. Thus

$$\sum_{h=-\infty}^{\infty} |c(h)| = \sigma^2 \sum_{h=-\infty}^{\infty} \left| \sum_j \psi_j \psi_{j+h} \right| \leq \sigma^2 \left(\sum_{j=-\infty}^{\infty} |\psi_j| \right)^2 < \infty.$$

then

$$|\mathrm{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}]| \leq \alpha(k_1)\alpha(k_2 - k_1)\alpha(k_3 - k_2) \quad (8.13)$$

where $\sum_{k=-\infty}^{\infty} \alpha(k) < \infty$.

(d) Often in proofs we use the assumption $\sum_r |c(r)| < \infty$. An analogous assumption for fourth order cumulants is $\sum_{k_1,k_2,k_3} |\kappa_4(k_1, k_2, k_3)| < \infty$. Based on the inequality (8.13), this assumption is often reasonable (such assumptions are often called Brillinger-type mixing conditions).

Point (c) and (d) are very important in the derivation of sampling properties of an estimator.

Example 8.2.1 - *We illustrate (d) for the causal AR(1) model $X_t = \phi X_{t-1} + \varepsilon_t$ (where $\{\varepsilon_t\}$ are iid random variables with finite fourth order cumulant $\kappa_4 = \mathrm{cum}(\varepsilon_t, \varepsilon_t, \varepsilon_t, \varepsilon_t)$). By using the MA($\infty$) representation $\sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$ (assuming $0 \leq k_1 \leq k_2 \leq k_3$) we have*

$$\begin{aligned} cum[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}] &= \sum_{j_0,j_1,j_2,j_3=0}^{\infty} \phi^{j_0+j_1+j_2+j_3} \mathrm{cum}\left[\varepsilon_{t-j_0}, \varepsilon_{t+k_1-j_1}, \varepsilon_{t+k_2-j_2}, \varepsilon_{t+k_3-j_3}\right] \\ &= \kappa_4 \sum_{j=0}^{\infty} \phi^j \phi^{j+k_1} \phi^{j+k_2} \phi^{j+k_3} = \kappa_4 \frac{\phi^{k_1+k_2+k_3}}{1-\phi^4}. \end{aligned}$$

The fourth order dependence decays as the lag increases. And this rate of decay is faster than the general bound $|\mathrm{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}]| \leq \alpha(k_1)\alpha(k_2 - k_1)\alpha(k_3 - k_2)$.

- *If $\{X_t\}_t$ are martingale differences and t_j are all different, then using (8.12) (the expansion of the fourth order cumulant in terms of moments) we have*

$$cum[X_{t_1}, X_{t_2}, X_{t_3}, X_{t_4}] = 0.$$

Remark 8.2.3 (Cumulants and dependence measures) *The summability of cumulants can be shown under various mixing and dependent type conditions. We mention a few below.*

- *Conditions for summability of cumulants for mixing processes are given in Statulevicius and Jakimavicius (1988) and Lahiri (2003).*
- *Conditions for summability of cumulants for physical dependence processes are given in Shao and Wu (2007), Theorem 4.1.*

Proof of equation (8.8) in Theorem 8.2.2

Our aim is to show

$$\mathrm{var}\left[\sqrt{n}\begin{pmatrix}\widehat{c}_n(0)\\ \widehat{c}_n(r_1)\\ \vdots\\ \widehat{c}_n(r_d)\end{pmatrix}\right] \to V_{d+1}$$

where

$$\begin{aligned}(V_{d+1})_{i,j} &= \sum_{k=-\infty}^{\infty} c(k)c(k+r_{i-1}-r_{j-1}) + \sum_{k=-\infty}^{\infty} c(k+r_{i-1}-1)c(k-r_{j-1}-1) + \\ &\quad \sum_{k=-\infty}^{\infty} \kappa_4(r_{i-1}-1,k,k+r_{j-1}-1). \qquad (8.14)\end{aligned}$$

To simplify notation we start by considering the variance

$$\mathrm{var}[\sqrt{n}\widehat{c}_n(r)] = \frac{1}{n}\sum_{t,\tau=1}^{n-|r|}\mathrm{cov}(X_tX_{t+r}, X_\tau X_{\tau+r}).$$

To prove the result, we use the identity (8.11); if A, B, C and D are mean zero random variables, then $\mathrm{cov}[AB, CD] = \mathrm{cov}[A,C]\mathrm{cov}[B,D] + \mathrm{cov}[A,D]\mathrm{cov}[B,C] + \mathrm{cum}[A,B,C,D]$. Using this identity we have

$$\begin{aligned}&\mathrm{var}[\widehat{c}_n(r)]\\ =\ & \frac{1}{n^2}\sum_{t,\tau=1}^{n-|r|}\Big(\underbrace{\mathrm{cov}(X_t,X_\tau)}_{=c(t-\tau)}\mathrm{cov}(X_{t+r},X_{\tau+r}) + \mathrm{cov}(X_t,X_{\tau+r})\mathrm{cov}(X_{t+r},X_\tau) + \underbrace{\mathrm{cum}(X_t,X_{t+r},X_\tau,X_{\tau+r})}_{\kappa_4(r,\tau-t,t+r-\tau)}\Big)\\ =\ & \frac{1}{n}\sum_{t,\tau=1}^{n-|r|}c(t-\tau)^2 + \frac{1}{n}\sum_{t,\tau=1}^{n-|r|}c(t-\tau-r)c(t+r-\tau) + \frac{1}{n}\sum_{t,\tau=1}^{n-|r|}k_4(r,\tau-t,\tau+r-t)\\ :=\ & I_n + II_n + III_n,\end{aligned}$$

where the above is due to strict stationarity of the time series. The benefit of using a cumulant expansion rather than a moment expansion is now apparent. Since cumulants act like a covariances, they do decay as the time gaps grow. This allows us to analysis each term I_n, II_n and III_n individually. This simplifies the analysis.

We first consider I_n. Either (i) by changing variables and letting $k = t - \tau$ and thus changing the limits of the summand in an appropriate way or (ii) observing that $\sum_{t,\tau=1}^{n-|r|} c(t-\tau)^2$ is the sum of the elements in the Toeplitz matrix

$$\begin{pmatrix} c(0)^2 & c(1)^2 & \ldots & c(n-1)^2 \\ c(-1)^2 & c(0)^2 & \ldots & c(n-2)^2 \\ \vdots & \vdots & \ddots & \vdots \\ c((n-1))^2 & c((n-2))^2 & \ldots & c(0)^2 \end{pmatrix},$$

(noting that $c(-k) = c(k)$) the sum I can be written as

$$I_n = \frac{1}{n}\sum_{t,\tau=1}^{n-|r|} c(t-\tau)^2 = \frac{1}{n}\sum_{k=-(n-1)}^{(n-1)} c(k)^2 \sum_{t=1}^{n-|k|} 1 = \sum_{k=-(n-1)}^{n-1} \left(\frac{n-|k|}{n}\right) c(k)^2.$$

To obtain the limit of the above we use dominated convergence. Precisely, since for all k, $(1 - |k|/n)c(k)^2 \to c(k)^2$ and $|\sum_{k=-(n-|r|)}^{n-|r|}(1-|k|/n)c(k)^2| \leq \sum_{k\in\mathbb{Z}} c(k)^2 < \infty$, by dominated convergence $I_n = \sum_{k=-(n-1)}^{n-1}(1-|k|/n)c(k)^2 \to \sum_{k=-\infty}^{\infty} c(k)^2$. Using a similar argument we can show that

$$\lim_{n\to\infty} II_n = \sum_{k=-\infty}^{\infty} c(k+r)c(k-r).$$

To derive the limit of III_n, we change variables $k = \tau - t$ to give

$$III_n = \sum_{k=-(n-|r|)}^{n-|r|} \left(\frac{n-|r|-|k|}{n}\right) k_4(r,k,k+r).$$

Again we use dominated convergence. Precisely, for all k, $(1-|k|/n)k_4(r,k,k+r) \to k_4(r,k,k+r)$ and $|\sum_{k=-(n-|r|)}^{n-|r|}(1-|k|/n)k_4(r,k,k+r)| \leq \sum_{k\in\mathbb{Z}} |k_4(r,k,k+r)| < \infty$ (by assumption). Thus by dominated convergence we have $III_n = \sum_{k=-(n-|r|)}^{n}(1-|k|/n)k_4(r,k,k+r) \to \sum_{k=-\infty}^{\infty} k_4(r,k,k+r)$. Altogether the limits of I_n, II_n and III_n give

$$\lim_{n\to\infty} \mathrm{var}[\sqrt{n}\widehat{c}_n(r)] = \sum_{k=-\infty}^{\infty} c(k)^2 + \sum_{k=-\infty}^{\infty} c(k+r)c(k-r) + \sum_{k=-\infty}^{\infty} \kappa_4(r,k,k+r).$$

Using similar set of arguments we obtain

$$\begin{aligned}
& \lim_{n\to\infty} \mathrm{cov}[\sqrt{n}\widehat{c}_n(r_1), \sqrt{n}\widehat{c}_n(r_2)] \\
\to\ & \sum_{k=-\infty}^{\infty} c(k)c(k+r_1-r_2) + \sum_{k=-\infty}^{\infty} c(k-r_1)c(k+r_2) + \sum_{k=-\infty}^{\infty} \kappa_4(r_1,k,k+r_2).
\end{aligned}$$

This result gives the required variance matrix V_{d+1} in Theorem 8.2.2.

Below, we show that under linearity the fourth order cumulant term has a simpler form. We will show

$$\sqrt{n}\begin{pmatrix} \widehat{\rho}_n(r_1) - \rho(r_1) \\ \vdots \\ \widehat{\rho}_n(r_d) - \rho(r_d) \end{pmatrix} \xrightarrow{\mathcal{P}} \mathcal{N}\left(0, GC_{d+1}G'\right).$$

We have already shown that in the general case the limit distribution of the sample correlations is $G(C_{d+1}+K_{d+1})G'$. Thus our objective here is to show that for linear time series the fourth order cumulant term is $GK_{d+1}G' = 0$.

Proof of Theorem 8.2.3 and the case of the vanishing fourth order cumulant

So far we have not used the structure of the time series to derive an expression for the variance of the sample covariance. However, to prove $GK_{d+1}G' = 0$ we require an explicit expression for K_{d+1}. The following result only holds for linear, univariate time series. We recall that

$$(K_{d+1})_{i,j} = \sum_{k=-\infty}^{\infty} \kappa_4(r_{i-1}, k, k+r_{j-1}).$$

By definition $\kappa_4(r_{i-1}, k, k+r_{j-1}) = \mathrm{cum}(X_0, X_{r_{i-1}}, X_k, X_{k+r_{j-1}})$. Further, we consider the specific case that X_t is a linear time series, where

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

$\sum_j |\psi_j| < \infty$, $\{\varepsilon_t\}$ are iid, $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{var}(\varepsilon_t) = \sigma^2$ and $\kappa_4 = \mathrm{cum}_4(\varepsilon_t)$. To find an expression for $(K_{d+1})_{i,j}$, consider the general sum

$$\begin{aligned} & \sum_{k=-\infty}^{\infty} \mathrm{cum}(X_0, X_{r_1}, X_k, X_{k+r_2}) \\ = & \sum_{k=-\infty}^{\infty} \mathrm{cum}\left(\sum_{j_1=-\infty}^{\infty} \psi_{j_1}\varepsilon_{-j_1}, \sum_{j_2=-\infty}^{\infty} \psi_{j_2}\varepsilon_{r_1-j_2}, \sum_{j_3=-\infty}^{\infty} \psi_{j_3}\varepsilon_{k-j_3}, \sum_{j_4=-\infty}^{\infty} \psi_{j_4}\varepsilon_{k+r_2-j_1}\right) \\ = & \sum_{k=-\infty}^{\infty} \sum_{j_1,\ldots,j_4=-\infty} \psi_{j_1}\psi_{j_2}\psi_{j_3}\psi_{j_4}\mathrm{cum}\left(\varepsilon_{-j_1}, \varepsilon_{r_1-j_2}, \varepsilon_{k-j_3}, \varepsilon_{k+r_2-j_1}\right). \end{aligned}$$

We recall from Section 8.2.3, if one of the variables above is independent of the other, then $\mathrm{cum}\left(\varepsilon_{-j_1}, \varepsilon_{r_1-j_2}, \varepsilon_{k-j_3}, \varepsilon_{k+r_2-j_1}\right) = 0$. This reduces the number of summands from five to two

$$\sum_{k=-\infty}^{\infty} \mathrm{cum}(X_0, X_{r_1}, X_k, X_{k+r_2}) \quad = \quad \kappa_4 \sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j-r_1} \psi_{j-k} \psi_{j-r_2-k}.$$

Changing variables $j_1 = j$ and $j_2 = j - k$ we have

$$\sum_{k=-\infty}^{\infty} \mathrm{cum}(X_0, X_{r_1}, X_k, X_{k+r_2}) = \kappa_4 \big(\sum_{j_1=-\infty}^{\infty} \psi_j \psi_{j-r_1}\big)\big(\sum_{j_2=-\infty}^{\infty} \psi_{j_2}\psi_{j_2-r_2}\big) = \kappa_4 \frac{c(r_1)}{\sigma^2}\frac{c(r_2)}{\sigma^2} = \frac{\kappa_4}{\sigma^4} c(r_1)c(r_2),$$

recalling that $\mathrm{cov}(X_t, X_{t+r}) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+r}$. Thus for linear time series

$$(K_{d+1})_{i,j} \quad = \quad \sum_{k=-\infty}^{\infty} \kappa_4(r_{i-1}, k, k + r_{j-1}) = \frac{\kappa_4}{\sigma^2} c(r_{i-1})c(j_{i-1})$$

and the matrix K_{d+1} is

$$K_{d+1} = \frac{\kappa_4}{\sigma^4} \underline{c}_{d+1}\underline{c}_{d+1}'$$

where $\underline{c}_{d+1}' = (c(0), c(r_1), \ldots, c(r_d))$. Substituting this representation of K_{d+1} into $GK_{d+1}G'$ gives

$$GK_{d+1}G' = \frac{\kappa_4}{\sigma^4} G\underline{c}_{d+1}\underline{c}_{d+1}'G'.$$

We recall from Remark 8.2.1 that G is a $d \times (d + 1)$ dimension matrix with null space $\underline{c}_{d+1}$. This immediately gives $G\underline{c}_{d+1} = 0$ and the result.

Exercise 8.1 *Under the assumption that $\{X_t\}$ are iid random variables show that $\hat{c}_n(1)$ is asymptotically normal.*

Hint: Let $m = n/(B+1)$ and partition the sum $\sum_{k=1}^{n-1} X_t X_{t+1}$ as follows

$$\begin{aligned}
\sum_{t=1}^{n-1} X_t X_{t+1} &= \sum_{t=1}^{B} X_t X_{t+1} + X_{B+1}X_{B+2} + \sum_{t=B+2}^{2B+1} X_t X_{t+1} + X_{2B+2}X_{2B+3} + \\
&\quad \sum_{t=2B+3}^{3B+2} X_t X_{t+1} + X_{3B+3}X_{3B+4} + \sum_{t=3B+4}^{4B+3} X_t X_{t+1} + \ldots \\
&= \sum_{j=0}^{m-1} U_{m,j} + \sum_{j=0}^{m-1} X_{(j+1)(B+1)} X_{(j+1)(B+1)+1}
\end{aligned}$$

where $U_{m,j} = \sum_{t=j(B+1)+1}^{j(B+1)+B} X_t X_{t+1}$. Show that the second term in the above summand is asymptotically negligible and show that the classical CLT for triangular arrays can be applied to the first term.

Exercise 8.2 *Under the assumption that $\{X_t\}$ is a MA(1) process, show that $\widehat{c}_n(1)$ is asymptotically normal.*

Exercise 8.3 *The block bootstrap scheme is a commonly used method for estimating the finite sample distribution of a statistic (which includes its variance). The aim in this exercise is to see how well the bootstrap variance approximates the finite sample variance of a statistic.*

(i) *In* `R` *write a function to calculate the autocovariance $\widehat{c}_n(1) = \frac{1}{n}\sum_{t=1}^{n-1} X_t X_{t+1}$.*

   Remember the function is defined as `cov1 = function(x){...}`

(ii) *Load the library boot* `library("boot")` *into* `R`. *We will use the block bootstrap, which partitions the data into blocks of lengths l and then samples from the blocks n/l times to construct a new bootstrap time series of length n. For each bootstrap time series the covariance is evaluated and this is done R times. The variance is calculated based on these R bootstrap estimates.*

   You will need to use the function `tsboot(tseries,statistic,R=100,l=20,sim="fixed")`. *tseries refers to the original data, statistic to the function you wrote in part (i) (which should only be a function of the data), R=is the number of bootstrap replications and l is the length of the block.*

Note that `tsboot(tseries,statistic,R=100,l=20,sim="fixed")$t` *will be vector of length* $R = 100$ *which will contain the bootstrap statistics, you can calculate the variance of this vector.*

(iii) Simulate the $AR(2)$ *time series* $arima.sim(list(order = c(2,0,0), ar = c(1.5, -0.75)), n = 128)$ *500 times. For each realisation calculate the sample autocovariance at lag one and also the bootstrap variance.*

(iv) Calculate the mean of the bootstrap variances and also the mean squared error (compared with the empirical variance), how does the bootstrap perform?

(iv) Play around with the bootstrap block length l*. Observe how the block length can influence the result.*

Remark 8.2.4 *The above would appear to be a nice trick, but there are two major factors that lead to the cancellation of the fourth order cumulant term*

- *Linearity of the time series*
- *Ratio between* $\hat{c}_n(r)$ *and* $\hat{c}_n(0)$*.*

Indeed this is not a chance result, in fact there is a logical reason why this result is true (and is true for many statistics, which have a similar form - commonly called ratio statistics). It is easiest explained in the Fourier domain. If the estimator can be written as

$$\frac{1}{n}\frac{\sum_{k=1}^{n}\phi(\omega_k)I_n(\omega_k)}{\frac{1}{n}\sum_{k=1}^{n}I_n(\omega_k)},$$

where $I_n(\omega)$ *is the periodogram, and* $\{X_t\}$ *is a linear time series, then we will show later that the asymptotic distribution of the above has a variance which is only in terms of the covariances* <u>*not*</u> *higher order cumulants. We prove this result in Section 11.5.*

8.3 Checking for correlation in a time series

Bartlett's formula if commonly used to check by 'eye; whether a time series is uncorrelated (there are more sensitive tests, but this one is often used to construct CI in for the sample autocovariances in several statistical packages). This is an important problem, for many reasons:

- Given a data set, we need to check whether there is dependence, if there is we need to analyse it in a different way.

- Suppose we fit a linear regression to time series data. We may to check whether the residuals are actually uncorrelated, else the standard errors based on the assumption of uncorrelatedness would be unreliable.

- We need to check whether a time series model is the appropriate model. To do this we fit the model and estimate the residuals. If the residuals appear to be uncorrelated it would seem likely that the model is correct. If they are correlated, then the model is inappropriate. For example, we may fit an AR(1) to the data, estimate the residuals ε_t, if there is still correlation in the residuals, then the AR(1) was not the correct model, since $X_t - \hat{\phi}X_{t-1}$ is still correlated (which it would not be, if it were the correct model).

We now apply Theorem 8.2.3 to the case that the time series are iid random variables. Suppose $\{X_t\}$ are iid random variables, then it is clear that it is trivial example of a (not necessarily Gaussian) linear process. We use (8.3) as an estimator of the autocovariances.

To derive the asymptotic variance of $\{\hat{c}_n(r)\}$, we recall that if $\{X_t\}$ are iid then $\rho(k) = 0$ for $k \neq 0$. Then by using Bartlett's formula we have

$$(W_h)_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

In other words, $\sqrt{n}\hat{\boldsymbol{\rho}}_n \overset{\mathcal{D}}{\to} \mathcal{N}(0, I_h)$. Hence the sample autocovariances at different lags are asymptotically uncorrelated and have variance one. This allows us to easily construct error bars for the sample autocovariances under the assumption of independence. If the vast majority of the sample autocovariance lie inside the error bars there is not enough evidence to suggest that the data is a realisation of a iid random variables (often called a white noise process). An example of the empirical ACF and error bars is given in Figure 8.1. We see that the empirical autocorrelations of the realisation from iid random variables all lie within the error bars. In contrast in Figure 8.2 we give a plot of the sample ACF of an AR(2). We observe that a large number of the sample autocorrelations lie outside the error bars.

Of course, simply checking by eye means that we risk misconstruing a sample coefficient that lies outside the error bars as meaning that the time series is correlated, whereas this could simply

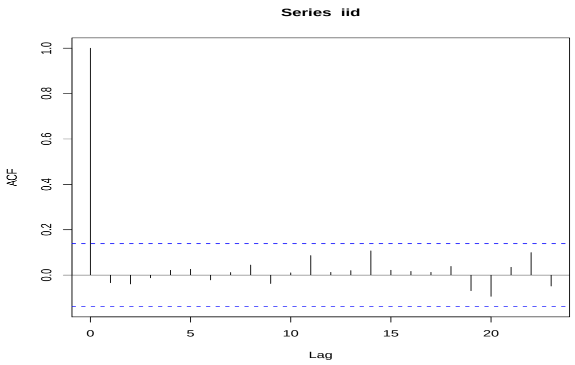


Figure 8.1: The sample ACF of an iid sample with error bars (sample size $n = 200$).

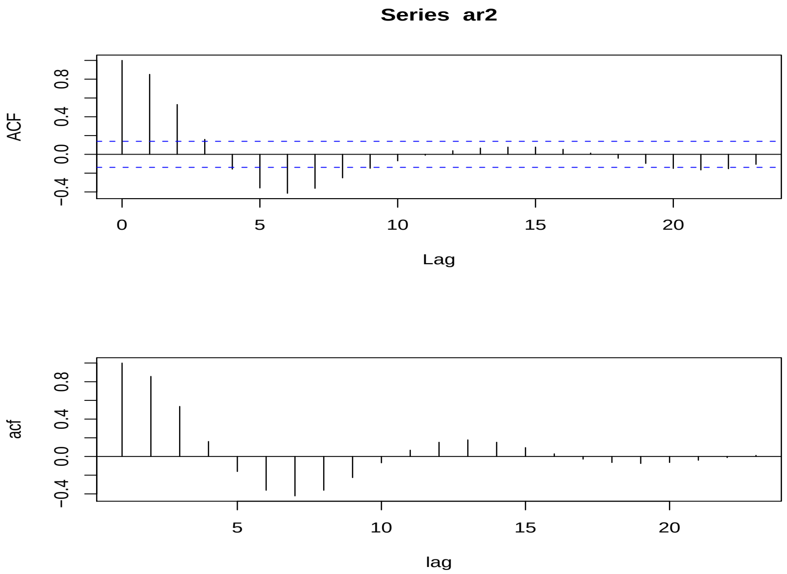


Figure 8.2: Top: The sample ACF of the AR(2) process $X_t = 1.5X_{t-1} + 0.75X_{t-2} + \varepsilon_t$ with error bars $n = 200$. Bottom: The true ACF.

be a false positive (due to multiple testing). To counter this problem, we construct a test statistic for testing uncorrelatedness. We test the hypothesis $H_0 : c(r) = 0$ for all r against H_A : at least one $c(r) \neq 0$.

A popular method for measuring correlation is to use the squares of the sample correlations

$$\mathcal{S}_h = n \sum_{r=1}^{h} |\widehat{\rho}_n(r)|^2. \tag{8.15}$$

Since under the null $\sqrt{n}(\hat{\rho}_n(h) - \rho(h)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I)$, under the null $\mathcal{S}_h$ asymptotically will have a χ^2-distribution with h degrees of freedom, under the alternative it will be a non-central (generalised) chi-squared. The non-centrality is what makes us reject the null if the alternative of correlatedness is true. This is known as the Box-Pierce (or Portmanteau) test. The Ljung-Box test is a variant on the Box-Pierce test and is defined as

$$\mathcal{S}_h = n(n+2) \sum_{r=1}^{h} \frac{|\hat{\rho}_n(r)|^2}{n-r}. \tag{8.16}$$

Again under the null of no correlation, asymptotically, $\mathcal{S}_h \xrightarrow{\mathcal{D}} \chi^2_h$. Generally, the Ljung-Box test is suppose to give more reliable results than the Box-Pierce test.

Of course, one needs to select h. In general, we do not have to use large h since most correlations will arise when the lag is small, However the choice of h will have an influence on power. If h is too large the test will loose power (since the mean of the chi-squared grows as $h \to \infty$), on the other hand choosing h too small may mean that certain correlations at higher lags are missed. How to selection h is discussed in several papers, see for example Escanciano and Lobato (2009).

Remark 8.3.1 (Do's and Don't of the Box-Jenkins or Ljung-Box test) *There is temptation to estimate the residuals from a model and test for correlation in the estimated residuals.*

- *Example 1* $Y_t = \sum_{j=1}^{p} \alpha_j x_{j,t} + \varepsilon_t$. *Suppose we want to know if the errors* $\{\varepsilon_t\}_t$ *are correlated. We test* H_0 : *errors are uncorrelated vs* H_A : *errors are correlated.*

  Suppose H_0 *is true.* $\{\varepsilon_t\}$ *are unobserved, but they can be estimated from the data. Then on the estimated residuals* $\{\widehat{\varepsilon}_t\}_t$ *we can test for correlation. We estimate the correlation based* *on the estimated residuals* $\widetilde{\rho}(r) = \widetilde{c}_n(r)/\widetilde{c}_n(0)$*, where*

$$\widetilde{c}_n(r) = \frac{1}{n}\sum_{t=1}^{n-|r|} \widehat{\varepsilon}_t\widehat{\varepsilon}_{t+r}.$$

It can be shown that $\sqrt{n}\widetilde{\rho}_n(r) \sim N(0,1)$ *and the Box-Jenkins or Ljung-Box test can be used. I.e.* $S_h \sim \chi^2_h$ *even when using the estimated residuals.*

- *Example 2 This example is a word of warning. Suppose* $Y_t = \phi Y_{t-1} + \varepsilon_t$*. We want to test* H_0 *:errors are uncorrelated vs* H_A *: errors are uncorrelated.*

  Suppose H_0 *is true.* $\{\varepsilon_t\}$ *are unobserved, but they can be estimated from the data. We estimate the correlation based on the estimated residuals (*$\widehat{\varepsilon}_t = Y_t - \widehat{\phi}Y_{t-1}$*),* $\widetilde{\rho}(r) = \widetilde{c}_n(r)/\widetilde{c}_n(0)$*, where*

$$\widetilde{c}_n(r) = \frac{1}{n}\sum_{t=1}^{n-|r|} \widehat{\varepsilon}_t\widehat{\varepsilon}_{t+r}.$$

  $\widetilde{\rho}_n(r)$ *is estimating zero* **but** $\sqrt{n}\widetilde{\rho}_n(r)$ *is not a standard normal. Thus* S_h *does not follow a standard chi-square distribution. This means the estimated residuals cannot be used to check for uncorrelatedness.*

  To understand the difference between the two examples see Section 8.6.

8.3.1 Relaxing the assumptions: The robust Portmanteau test (advanced)

One disadvantage of the Box-Pierce/Portmanteau test described above is that it requires under the null that the time series is *independent* not just uncorrelated. Even though the test statistic can only test for correlatedness and not dependence. As an illustration of this, in Figure **??** we give the QQplot of the $\mathcal{S}_2$ (using an ARCH process as the time series) against a chi-square distribution. We recall that despite the null being true, the test statistic deviates considerably from a chi-square. For this time series, we would have too many false positive despite the time series being uncorrelated. Thus the Box-Pierce test only gives reliable results for linear time series.

In general, under the null of no correlation we have

$$\mathrm{cov}\left(\sqrt{n}\widehat{c}_n(r_1), \sqrt{n}\widehat{c}_n(r_2)\right) = \begin{cases} \sum_k \kappa_4(r_1,k,k+r_2) & r_1 \neq r_2 \\ c(0)^2 + \sum_k \kappa_4(r,k,k+r) & r_1 = r_2 = (r) \end{cases}$$

Thus despite $\widehat{c}_n(r)$ being asymptotically normal we have

$$\sqrt{n}\frac{\widehat{c}_n(r)}{c(0)} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, 1 + GK_2G'\right),$$

where the cumulant term GK_2G tends to be positive. This results in the Box-Pierce test underestimating the variance, and the true quantiles of $\mathcal{S}_2$ (see Figure **??**) being larger than the chi square quantiles.

However, there is an important subset of uncorrelated time series, which are dependent, where a slight modification of the Box-Pierce test does give reliable results. This subset includes the aforementioned ARCH process and is a very useful test in financial applications. As mentioned in (**??**) ARCH and GARCH processes are uncorrelated time series which are martingale differences. We now describe the robust Portmanteau test, which is popular in econometrics as it is allows for uncorrelated time series which are martingale differences and an additional joint moment condition which we specify below (so long as it is stationary and its fourth moment exists).

We recall that $\{X_t\}_t$ is a martingale difference if

$$\mathrm{E}(X_t|X_{t-1}, X_{t-2}, X_{t-3}, \ldots) = 0.$$

Martingale differences include independent random variables as a special case. Clearly, from this definition $\{X_t\}$ is uncorrelated since for $r > 0$ and by using the definition of a martingale difference we have

$$\begin{aligned} \mathrm{cov}(X_t, X_{t+r}) &= \mathrm{E}(X_tX_{t+r}) - \mathrm{E}(X_t)\mathrm{E}(X_{t+r}) \\ &= \mathrm{E}(X_t\mathrm{E}(X_{t+r}|X_t)) - \mathrm{E}(\mathrm{E}(X_t|X_{t-1}))\mathrm{E}(\mathrm{E}(X_{t+r}|X_{t+r-1})) = 0. \end{aligned}$$

Thus a martingale difference sequence is an uncorrelated sequence. However, martingale differences have more structure than uncorrelated random variables, thus allow more flexibility. For a test to be simple we would like that the sample covariance between different lags is asymptotically zero.

This can be achieved for martinagle differences plus an important *additional* condition:

$$\mathrm{E}[X_t^2 X_{s_1} X_{s_2}] = 0 \qquad t > s_1 \neq s_2. \tag{8.17}$$

To understand why, consider the sample covariance

$$\mathrm{cov}\left(\sqrt{n}\widehat{c}_n(r_1), \sqrt{n}\widehat{c}_n(r_2)\right) = \frac{1}{n}\sum_{t_1,t_2} \mathrm{cov}\left(X_{t_1}X_{t_1+r_1}, X_{t_2}X_{t_2+r_2}\right)$$

Under the null, the above is

$$\mathrm{cov}\left(\sqrt{n}\widehat{c}_n(r_1), \sqrt{n}\widehat{c}_n(r_2)\right) = \frac{1}{n}\sum_{t_1,t_2} \mathrm{E}\left(X_{t_1}X_{t_1+r_1}X_{t_2}X_{t_2+r_2}\right).$$

We show that under the null hypothesis, many of the above terms are zero (when $r_1 \neq r_2$), however there are some exceptions, which require the additional moment condition.

For example, if $t_1 \neq t_2$ and suppose for simplicity $t_2 + r_2 > t_2, t_1, t_1 + r_1$. Then

$$\mathrm{E}(X_{t_1}X_{t_1+r_1}X_{t_2}X_{t_2+r_2}) = \mathrm{E}\left(X_{t_1}X_{t_1+r_2}X_{t_2}\mathrm{E}(X_{t_2+r_2}|X_{t_1}, X_{t_1+r_2}, X_{t_2})\right) = 0 \tag{8.18}$$

and if $r_1 \neq r_2$ (assume $r_2 > r_1$) by the same argument

$$\mathrm{E}(X_tX_{t+r_1}X_tX_{t+r_2}) = \mathrm{E}\left(X_t^2X_{t+r_1}\mathrm{E}(X_{t+r_2}| \underbrace{X_t^2, X_{t+r_1}}_{\subset\sigma(X_{t+r_2-1}, X_{t+r_2-2},\ldots)})\right) = 0.$$

However, in the case that $t_1 + r_1 = t_2 + r_2$ ($r_1 \neq r_2 \geq 0$, since $r_1 \neq r_2$, then this implies $t_1 \neq t_2$) we have

$$\mathrm{E}(X_{t_1+r_1}^2 X_{t_1}X_{t_2}) \neq 0,$$

even when X_t are martingale arguments. Consequently, we do not have that $\mathrm{cov}(X_{t_1}X_{t_1+r_1}, X_{t_2}X_{t_2+r_2}) = 0$. However, by including the additional moment condition that $\mathrm{E}[X_t^2X_{s_1}X_{s_2}] = 0$ for $t > s_1, \neq s_2$, then we have $\mathrm{cov}(X_{t_1}X_{t_1+r_1}, X_{t_2}X_{t_2+r_2}) = 0$ for all t_1 and t_2 when $r_1 \neq r_2$.

The above results can be used to show that the variance of $\widehat{c}_n(r)$ (under the assumption that the time series martingale differences and $\mathrm{E}[X_t^2 X_{s_1} X_{s_2}] = 0$ for $t > s_1, s_2$) has a very simple form

$$
\begin{aligned}
\operatorname{var}\left(\sqrt{n}\widehat{c}_n(r)\right) &= \frac{1}{n}\sum_{t_1,t_2=1}^{n} \operatorname{cov}\left(X_{t_1}X_{t_1+r}, X_{t_2}X_{t_2+r}\right) \\
&= \frac{1}{n}\sum_{t_1,t_2=1}^{n} \mathrm{E}\left(X_{t_1}X_{t_1+r}X_{t_2}X_{t_2+r}\right) = \frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left(X_t^2 X_{t+r}^2\right) = \underbrace{\mathrm{E}(X_0^2 X_r^2)}_{\text{by stationarity}}
\end{aligned}
$$

and if $r_1 \neq r_2$ then $\operatorname{cov}(\widehat{c}_n(r_1), \widehat{c}_n(r_2)) = 0$. Let $\sigma_r^2 = \mathrm{E}\left(X_0^2 X_r^2\right)$. Then we have that under the null hypothesis (and suitable conditions to ensure normality) that

$$
\sqrt{n}\begin{pmatrix} \widehat{c}_n(1)/\sigma_1 \\ \widehat{c}_n(2)/\sigma_2 \\ \vdots \\ \widehat{c}_n(h)/\sigma_h \end{pmatrix} \overset{\mathcal{D}}{\rightarrow} \mathcal{N}\left(0, I_h\right).
$$

It is straightforward to estimate the σ_r^2 with

$$
\widehat{\sigma}_r^2 = \frac{1}{n}\sum_{t=1}^{n} X_t^2 X_{t+r}^2.
$$

Thus a similar squared distance as the Box-Pierce test is used to define the Robust Portmanteau test, which is defined as

$$
\mathcal{R}_h = n\sum_{r=1}^{h}\frac{|\widehat{c}_n(r)|^2}{\widehat{\sigma}_r^2}.
$$

Under the null hypothesis (assuming stationarity and martingale differences) asymptotically $\mathcal{R}_h \overset{\mathcal{D}}{\rightarrow} \chi_h^2$ (for h kept fixed).

Remark 8.3.2 (ARCH and the Robust Portmanteau test) *If I remember correctly the reason the above condition holds for ARCH models is (we assume wlog $s_2 > s_1$)*

$$
\begin{aligned}
\mathrm{E}[X_t^2 X_{s_1} X_{s_2}] &= \mathrm{E}[\eta_t^2]\mathrm{E}[\sigma_t^2 \sigma_{s_1}\sigma_{s_1}\eta_{s_2}\eta_{s_1}] \\
&= \mathrm{E}[\eta_t^2]\mathrm{E}[\eta_{s_2}\eta_{s_1}\mathrm{E}[\sigma_t^2\sigma_{s_2}\sigma_{s_1}|\mathcal{F}_{s_1-1}]] \\
&= \mathrm{E}[\eta_{s_1}]\mathrm{E}[\eta_{s_2}]\mathrm{E}[\sigma_t^2\sigma_{s_2}\sigma_{s_1}|\mathcal{F}_{s_1-1}]] = 0,
\end{aligned}
$$

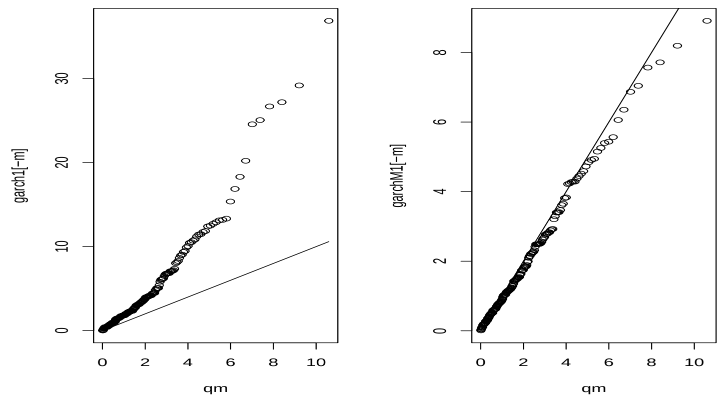


Figure 8.3: Using ARCH(1) time series over 200 replications Left: $\mathcal{S}_2$ against the quantiles of a chi-square distribution with 2df for an ARCH process. Right: $\mathcal{R}_2$ against the quantiles of a chi-square distribution with 2df for an ARCH process.

To see how this test performs, in the right hand plot in Figure 8.3 we give the quantile quantile plot of $\mathcal{R}_h$ against the chi-squared distribution. We observe that it lies pretty much on the $x = y$ line. Moreover, the test results at the 5% level are given in Table 8.1. We observe that it is close to the stated 5% level and performs far better than the classical Box-Pierce test.

ARCH Box-Pierce	26%
ARCH Robust Portmanteau	4.5%

Table 8.1: Proportion of rejections under the null hypothesis. Test done at the 5% level over 200 replications.

The robust Portmanteau test is a useful generalisation of the Box-Pierce test, however it still requires that the time series under the null satisfies the martingale difference property and the moment condition. These conditions cannot be verified. Consider for example the uncorrelated time series

$$X_{t+1} = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} - \frac{\phi}{1-\phi^2} \varepsilon_{t+1}$$

where $\{\varepsilon_t\}$ are uncorrelated random variables from the ARCH process $\varepsilon_t = Z_t \sigma_t$ and $\sigma_t^2 = a_0 + a_1 \varepsilon_{t-1}^2$. Despite ε_t being martingale differences, X_t are not martingale differences. Thus the robust Portmanteau test will not necessarily give satisfactory results for this uncorrelated time series.

Methods have been developed for these general time series methods, including:

- The robust test for white noise proposed in Dalla et al. (2019).
- Bootstrap methods. These include the block bootstrap (Künsch (1989), Liu and Singh (1992) and Lahiri (2003)), the stationary bootstrap (Politis and Romano (1994)), the sieve bootstrap (Kreiss (1992) and Kreiss et al. (2011)) and the spectral bootstrap (Hurvich and Zeger (1987), Franke and Härdle (1992), Dahlhaus and Janas (1996) and Dette and Paparoditis (2009)). Please keep in mind that this is an incomplete list.
- Estimating the variance of the sample covariance using spectral methods or long-run variance methods (together with fixed-b asymptotics have been used to obtain a more reliable finite sample estimator of the distribution).

Finally a few remarks about ACF plots in general

- It is clear that the theoretical autocorrelation function of an MA(q) process is such that $\rho(r) = 0$ if $|r| > q$. Thus from the theoretical ACF we can determine the order of the process. By a similar argument the variance matrix of an MA(q) will be bandlimited, where the band is of order q.

  However, we *cannot* determine the order of an moving average process from the empirical ACF plot. The critical values seen in the plot only correspond to the case the process is iid, they cannot be used as a guide for determining order.
- Often a model is fitted to a time series and the residuals are evaluated. To see if the model was appropriate, and ACF plot of empirical correlations corresponding to the estimated residuals. Even if the true residuals are iid, the variance of the empirical residuals correlations will not be (**??**). Li (1992) shows that the variance depends on the sampling properties of the model estimator.
- Misspecification, when the time series contains a time-dependent trend.

8.4 Checking for partial correlation

We recall that the partial correlation of a stationary time series at lag t is given by the last coefficient of the best linear predictor of X_{m+1} given $\{X_j\}_{j=1}^{m}$ i.e. ϕ_m where $\widehat{X}_{m+1|m} = \sum_{j=1}^{m} \phi_j X_{m+1-j}$. Thus

ϕ_m can be estimated using the Yule-Walker estimator or least squares (more of this later) and the sampling properties of the estimator are determined by the sampling properties of the estimator of an AR(m) process. We state these now. We assume $\{X_t\}$ is a AR(p) time series of the form

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2. Suppose an AR(m) model is fitted to the data using the Yule-Walker estimator, we denote this estimator as $\widehat{\boldsymbol{\phi}}_m = \widehat{\Sigma}_m^{-1}\underline{r}_m$. Let $\widehat{\phi}_m = (\widehat{\phi}_{m1}, \ldots, \widehat{\phi}_{mm})$, the estimator of the partial correlation at lag m is $\widehat{\phi}_{mm}$. Assume $m \geq p$. Then by using Theorem 9.2.1 (see also Theorem 8.1.2, Brockwell and Davis (1998)) we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\phi}}_m - \boldsymbol{\phi}_m\right) \xrightarrow{\mathcal{P}} N(0, \sigma^2 \Sigma_m^{-1}).$$

where $\boldsymbol{\phi}_m$ are the true parameters. If $m > p$, then $\boldsymbol{\phi}_m = (\phi_1, \ldots, \phi_p, 0, \ldots, 0)$ and the last coefficient has the marginal distribution

$$\sqrt{n}\widehat{\phi}_{mm} \xrightarrow{\mathcal{P}} N(0, \sigma^2 \Sigma^{mm}).$$

Since $m > p$, we can obtain a closed for expression for Σ^{mm}. By using Remark 6.3.1 we have $\Sigma^{mm} = \sigma^{-2}$, thus

$$\sqrt{n}\widehat{\phi}_{mm} \xrightarrow{\mathcal{P}} N(0, 1).$$

Therefore, for lags $m > p$ the partial correlations will be asymptotically pivotal. The errors bars in the partial correlations are $[-1.96n^{-1/2}, 1.96n^{-1/2}]$ and these can be used as a guide in determining the order of the autoregressive process (note there will be dependence between the partial correlation at different lags).

This is quite a surprising result and very different to the behaviour of the sample autocorrelation function of an MA(p) process.

Exercise 8.4

(a) Simulate a mean zero invertible MA(1) process (use Gaussian errors). Use a reasonable sample size (say $n = 200$). Evaluate the sample correlation at lag 2, $\widehat{rho}_n(2)$. Note the sample correlation

at lag two is estimating 0. Do this 500 times.

- *Calculate of proportion of sample covariances* $|\widehat{\rho}_n(2)| > 1.96/\sqrt{n}$
- *Make a QQplot of* $\widehat{\rho}_n(2)/\sqrt{n}$ *against a standard normal distribution. What do you observe?*

(b) Simulate a causal, stationary $AR(1)$ *process (use Gaussian errors). Use a reasonable sample size (say* $n = 200$*). Evaluate the sample partial correlation at lag 2,* $\widehat{\phi}_n(2)$*. Note the sample partial correlation at lag two is estimating 0. Do this 500 times.*

- *Calculate of proportion of sample partial correlations* $|\widehat{\phi}_n(2)| > 1.96/\sqrt{n}$
- *Make a QQplot of* $\widehat{\phi}_n(2)/\sqrt{n}$ *against a standard normal distribution. What do you observe?*

8.5 The Newey-West (HAC) estimator

In this section we focus on the estimation of the variance of

$$\widehat{\theta}_n = \frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t,$$

where $\{\mathbf{u}_t\}_t$ are deterministic regressors and $\{\varepsilon_t\}$ is a time series. Quantities of the form $\widehat{\theta}_n$ arise in several applications. One important application is in linear regression, which we summarize below.

In Section 3.2.1 we showed that the least squares estimator of the cofficients in

$$Y_t = \beta_0 + \sum_{j=1}^{p}\beta_j u_{t,j} + \varepsilon_t = \boldsymbol{\beta}'\mathbf{u}_t + \varepsilon_t,$$

is

$$\hat{\beta}_n = \arg\min \mathcal{L}_n(\boldsymbol{\beta}) = (\sum_{t=1}^{n}\mathbf{u}_t\mathbf{u}_t')^{-1}\sum_{t=1}^{n}Y_t\mathbf{u}_t.$$

The variance of $\hat{\beta}_n$ is derived using

$$\begin{aligned}\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right]'\sum_{t=1}^{n}\mathbf{u}_t\mathbf{u}_t' &= \sum_{t=1}^{n}\mathbf{u}_t'\varepsilon_t \\ \Rightarrow \left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right] &= \left(\sum_{t=1}^{n}\mathbf{u}_t\mathbf{u}_t'\right)^{-1}\sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t = \left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\mathbf{u}_t'\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t.\end{aligned}$$

Using this expression we have

$$\mathrm{var}\left[\hat{\boldsymbol{\beta}}_n\right] = \left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\mathbf{u}_t'\right)^{-1} \mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t\right)\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\mathbf{u}_t'\right)^{-1}.$$

Hence the variance of $\hat{\boldsymbol{\beta}}_n$ is based on var $\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t\right)$ which is

$$\begin{aligned}
\mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t\right) &= \frac{1}{n^2}\sum_{t,\tau=1}^{n}\mathrm{cov}[\varepsilon_t,\varepsilon_\tau]\mathbf{u}_t\mathbf{u}_\tau' \\
&= \frac{1}{n^2}\sum_{t=1}^{n}\mathrm{var}[\varepsilon_t]\mathbf{u}_t\mathbf{u}_t' + \frac{1}{n^2}\sum_{t=1}^{n}\sum_{\tau\neq t}^{n}\mathrm{cov}[\varepsilon_t,\varepsilon_\tau]\mathbf{u}_t\mathbf{u}_\tau' \\
&= \frac{1}{n^2}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathrm{cov}[\varepsilon_t,\varepsilon_\tau]\mathbf{u}_t\mathbf{u}_\tau'.
\end{aligned}$$

In the case of stationarity of $\{\varepsilon_t\}_t$, the above reduces

$$n\mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t\right) = \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}c(t-\tau)\mathbf{u}_t\mathbf{u}_\tau', \qquad (8.19)$$

where $c(t-\tau) = \mathrm{cov}[\varepsilon_t,\varepsilon_\tau]$.

We start by motivating the estimator of (8.20), we start with the special case that $\mathbf{u}_t = 1$ for all t. In this case (8.20) reduces to

$$n\mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t\right) = \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}c(t-\tau). \qquad (8.20)$$

Since $\mathrm{E}[\varepsilon_t\varepsilon_\tau] = c(t-\tau)$, as an estimator of the above we can potentially replace $c(t-\tau)$ with $\varepsilon_t\varepsilon_\tau$ to give the estimator

$$\widehat{\sigma}_{n,n}^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\varepsilon_t\varepsilon_\tau = \underbrace{\sum_{r=-n}^{n}\widehat{c}_r}_{\text{due to a change of variables}},$$

where $\widehat{c}_r = n^{-1}\sum_{t=1}^{n-|r|}\varepsilon_t\varepsilon_{t+r}$. We recall that in Section 8.2.1 we studied the sampling properties of $\widehat{c}_r$ and showed that $\mathrm{var}[\widehat{c}_r] = O(1/n)$. As $\frac{1}{n}\sum_{r=-n}^{n}\widehat{c}_r$ consists of the sum of all n sample covariances, this would suggest $\mathrm{var}[\widehat{\alpha}_n] = O(\sum_{r=1}^{n}n^{-1}) = O(1)$. Thus $\widehat{\alpha}_n$ is an inconsistent estimator of the variance. Calculations show that this is indeed the case. We discuss a very similar issue in Section

11.3 when estimating the spectral density function.

However, $\widehat{\sigma}^2_{n,n}$ suggests an alternative approach to estimation. As the autocovariance decays as the lag r grows, it is not necessary to estimate *all* the covariance and instead to truncate the number of covariances to be estimated i.e. use

$$\widehat{\sigma}^2_{m,n} = \sum_{r=-m}^{m} \lambda_m(r)\widehat{c}_r = \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{m} \lambda_m(t-\tau)\varepsilon_t\varepsilon_\tau, \tag{8.21}$$

where $\lambda_m(r)$ is a a so called lagged window which is zero for $|r| > m$. It can be shown that this truncation technique induces a bias in the estimation scheme (i.e. $\mathrm{E}[\widehat{\sigma}^2_{m,n}] \neq n\mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t\right)$) but the variance converges to zero ($\mathrm{E}[\widehat{\sigma}^2_{m,n}] = O(m/n)$). By balancing the bias and variance we can mind suitable choice of m such that $\widehat{\sigma}^2_{m,n}$ is a consistent estimator of σ^2.

The estimator $\widehat{\sigma}^2_{m,n}$ can be generalized to include the case $\mathbf{u}_t \neq 1$ and nonstationary errors $\{\varepsilon_t\}$. We recall that

$$n\mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{u}_t\varepsilon_t\right) = \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathrm{cov}[\varepsilon_t,\varepsilon_\tau]\mathbf{u}_t\mathbf{u}_\tau'. \tag{8.22}$$

We assume that $|\mathrm{cov}[\varepsilon_t,\varepsilon_\tau]| \leq |t-\tau|^{-\kappa}$ (where $\kappa > 1$). Since $\varepsilon_t\varepsilon_\tau$ can be treated as a "preestimator" (an initial estimator) of $\mathrm{cov}[\varepsilon_t,\varepsilon_\tau]$ we replace $\mathrm{cov}[\varepsilon_t,\varepsilon_\tau]$ in (8.22) with $\varepsilon_t\varepsilon_\tau$ and weight it with $\lambda_m(\cdot)$ to yield the Newey-West/HAC estimator

$$\widehat{\sigma}^2_{m,n} = \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\lambda_{m,n}(t-\tau)\varepsilon_t\varepsilon_\tau\mathbf{u}_t\mathbf{u}_\tau'. \tag{8.23}$$

Choices of weight functions are discussed in Section 11.3.1. The estimator (8.23) is closely related to spectral density estimation at frequency zero. The sampling properties of $\widehat{\sigma}^2_{m,n}$ are similar to those of spectral density estimation and can be found in (11.3.1). Further details can be found in Andrews (1990).

8.6 Checking for Goodness of fit (advanced)

To check for adequency of a model, after fitting a model to the data the sample correlation of the estimated residuals is evaluated. If there appears to be no correlation in the estimated residuals (so the residuals are near uncorrelated) then the model is determined to adequately fit the data.

Consider the general model

$$X_t = g(Y_t, \theta) + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid random variables and ε_t is independent of $Y_t, Y_{t-1}, \ldots$. Note Y_t can be a vector, such as $Y_{t-1} = (X_{t-1}, X_{t-2}, \ldots, X_{t-p})$ and examples of models which satisfy the above include the AR(p) process. We will assume that $\{X_t, Y_t\}$ is a stationary ergodic process. Further to simplify the discussion we will assume that θ is univariate, it is straightforward to generalize the discussion below to the multivariate case.

Let $\widehat{\theta}$ denote the least squares estimator of θ i.e.

$$\widehat{\theta} = \arg\min \sum_{t=1}^{n} \left(X_t - g(Y_t, \theta)\right)^2 . \qquad (8.24)$$

Using the "usual" Taylor expansion methods (and assuming all the usual conditions are satisfied, such as $|\widehat{\theta} - \theta| = O_p(n^{-1/2})$ etc) then it can be shown that

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) = \mathcal{I}^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t \frac{\partial g(Y_t, \theta)}{\partial \theta} + o_p(1) \text{ where } \mathcal{I} = \mathrm{E}\left(\frac{\partial g(Y_t, \theta)}{\partial \theta}\right)^2 .$$

$\{\varepsilon_t \frac{\partial g(Y_t,\theta)}{\partial \theta}\}$ are martingale differences, which is why $\sqrt{n}\left(\widehat{\theta} - \theta\right)$ is asymptotically normal, but more of this in the next chapter. Let $\mathcal{L}_n(\theta)$ denote the least squares criterion. Note that the above is true because

$$\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} = -2 \sum_{t=1}^{n} \left[X_t - g(Y_t, \theta)\right] \frac{\partial g(Y_t, \theta)}{\partial \theta}$$

and

$$\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2} = -2 \sum_{t=1}^{n} \left[X_t - g(Y_t, \theta)\right] \frac{\partial^2 g(Y_t, \theta)}{\partial \theta^2} + 2 \sum_{t=1}^{n} \left(\frac{\partial g(Y_t, \theta)}{\partial \theta}\right)^2 ,$$

thus at the true parameter, θ,

$$\frac{1}{n} \frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2} \xrightarrow{\mathcal{P}} 2\mathcal{I}.$$

Based on (8.24) we estimate the residuals using

$$\widehat{\varepsilon}_t = X_t - g(Y_t, \widehat{\theta})$$

and the sample correlation with $\widehat{\rho}(r) = \widehat{c}(r)/\widehat{c}(0)$ where

$$\widehat{c}(r) = \frac{1}{n}\sum_{t=1}^{n-|r|}\sum_{t} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t+r}.$$

Often it is (wrongly) assumed that one can simply apply the results in Section 8.3 when checking for adequacy of the model. That is make an ACF plot of $\widehat{\rho}(r)$ and use $[-n^{-1/2}, n^{1/2}]$ as the error bars. However, since the parameters have been estimated the size of the error bars will change. In particular, under the null that the model is correct we will show that

$$\sqrt{n}\widehat{\rho}(r) = \mathcal{N}\left(0, \underbrace{1}_{\text{iid part}} - \underbrace{\frac{\sigma^2}{c(0)}\mathcal{J}_r\mathcal{I}^{-1}\mathcal{J}_r}_{\text{due to parameter estimation}}\right)$$

where $c(0) = \mathrm{var}[X_t]$, $\sigma^2 = \mathrm{var}(\varepsilon_t)$ and $\mathcal{J}_r = \mathrm{E}[\frac{\partial g(Y_{t+r},\theta)}{\partial\theta}\varepsilon_t]$ and $\mathcal{I} = \mathrm{E}\left(\frac{\partial g(Y_t,\theta)}{\partial\theta}\right)^2$ (see, for example, Li (1992)). Thus the error bars under the null are

$$\left[\pm\left(\frac{1}{\sqrt{n}}\left[1 - \frac{\sigma^2}{c(0)}\mathcal{J}_r\mathcal{I}^{-1}\mathcal{J}_r\right]\right)\right].$$

Estimation of the parameters means the inclusion of the term $\frac{\sigma^2}{c(0)}\mathcal{J}_r\mathcal{I}^{-1}\mathcal{J}_r$. If the lag r is not too small then $\mathcal{J}_r$ will be close to zero and the $[\pm 1/\sqrt{n}]$ approximation is fine, but for small r, $\mathcal{J}_r\mathcal{I}^{-1}\mathcal{J}_r$ can be large and positive, thus the error bars, $\pm n^{-1/2}$, are too wide. Thus one needs to be a little cautious when interpreting the $\pm n^{-1/2}$ error bars. Note if there is no dependence between ε_t and Y_{t+r} then using the usual error bars is fine.

Remark 8.6.1 *The fact that the error bars get narrower after fitting a model to the data seems a little strange. However, it is far from unusual. One explanation is that the variance of the estimated residuals tend to be less than the true residuals (since the estimated residuals contain less information about the process than the true residuals). The most simplest example are iid observations $\{X_i\}_{i=1}^n$ with mean μ and variance σ^2. The variance of the "estimated residual" $X_i - \bar{X}$ is $(n-1)\sigma^2/n$.*

We now derive the above result (using lots of Taylor expansions). By making a Taylor expansion similar to (**??**) we have

$$\sqrt{n}\left[\widehat{\rho}_n(r)-\rho(r)\right]\sqrt{n}\frac{[\hat{c}_n(r)-c(r)]}{c(0)}-\sqrt{n}\left[\hat{c}_n(0)-c(0)\right]\frac{c(r)}{c(0)^2}+O_p(n^{-1/2}).$$

However, under the "null" that the correct model was fitted to the data we have $c(r)=0$ for $|r|>0$, this gives

$$\sqrt{n}\widehat{\rho}_n(r)=\sqrt{n}\frac{\widehat{c}_n(r)}{c(0)}+o_p(1),$$

thus the sampling properties of $\widehat{\rho}_n(r)$ are determined by $\widehat{c}_n(r)$, and we focus on this term. It is easy to see that

$$\sqrt{n}\widehat{c}_n(r)=\frac{1}{\sqrt{n}}\sum_{t=1}^{n-r}\left(\varepsilon_t+g(\theta,Y_t)-g(\widehat{\theta},Y_t)\right)\left(\varepsilon_{t+r}+g(\theta,Y_{t+r})-g(\widehat{\theta},Y_{t+r})\right).$$

Heuristically, by expanding the above, we can see that

$$\sqrt{n}\widehat{c}_n(r)\approx\frac{1}{\sqrt{n}}\sum_{t=1}^{n-r}\varepsilon_t\varepsilon_{t+r}+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_{t+r}\left(g(\theta,Y_t)-g(\widehat{\theta},Y_t)\right)+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t\left(g(\theta,Y_{t+r})-g(\widehat{\theta},Y_{t+r})\right),$$

then by making a Taylor expansion of $g(\widehat{\theta},\cdot)$ about $g(\theta,\cdot)$ (to take $(\widehat{\theta}-\theta)$ out of the sum)

$$\begin{aligned}\sqrt{n}\widehat{c}_n(r) &\approx \frac{1}{\sqrt{n}}\sum_{t=1}^{n-r}\varepsilon_t\varepsilon_{t+r}+\frac{(\widehat{\theta}-\theta)}{\sqrt{n}}\left[\sum_{t=1}^{n}\varepsilon_{t+r}\frac{\partial g(\theta,Y_t)}{\partial\theta}+\varepsilon_t\frac{\partial g(\theta,Y_{t+r})}{\partial\theta}\right]+o_p(1)\\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n-r}\varepsilon_t\varepsilon_{t+r}+\sqrt{n}(\widehat{\theta}-\theta)\frac{1}{n}\left[\sum_{t=1}^{n}\varepsilon_{t+r}\frac{\partial g(\theta,Y_t)}{\partial\theta}+\varepsilon_t\frac{\partial g(\theta,Y_{t+r})}{\partial\theta}\right]+o_p(1).\end{aligned}$$

We make this argument precise below. Making a Taylor expansion we have

$$\begin{aligned}\sqrt{n}\widehat{c}_n(r) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n-r}\left(\varepsilon_t-(\widehat{\theta}-\theta)\frac{\partial g(\theta,Y_t)}{\partial\theta}+\frac{(\widehat{\theta}-\theta)^2}{2}\frac{\partial^2 g(\bar{\theta}_t,Y_t)}{\partial\theta^2}\right)\times\\ &\quad\left(\varepsilon_{t+r}-(\widehat{\theta}-\theta)\frac{\partial g(\theta,Y_{t+r})}{\partial\theta}+\frac{(\widehat{\theta}-\theta)^2}{2}\frac{\partial^2 g(\bar{\theta}_{t+r},Y_{t+r})}{\partial\theta^2}\right)\\ &= \sqrt{n}\widetilde{c}_n(r)-\sqrt{n}(\widehat{\theta}-\theta)\frac{1}{n}\sum_{t=1}^{n-r}\left(\varepsilon_t\frac{\partial g(\theta,Y_{t+r})}{\partial\theta}+\varepsilon_{t+r}\frac{\partial g(\theta,Y_t)}{\partial\theta}\right)+O_p(n^{-1/2})(8.25)\end{aligned}$$

where θ_t lies between $\widehat{\theta}$ and θ and

$$\widetilde{c}_n(r) = \frac{1}{n}\sum_{t=1}^{n-r}\varepsilon_t\varepsilon_{t+r}.$$

We recall that by using ergodicity we have

$$\frac{1}{n}\sum_{t=1}^{n-r}\left(\varepsilon_t\frac{\partial g(\theta, Y_{t+r})}{\partial\theta} + \varepsilon_{t+r}\frac{\partial g(\theta, Y_t)}{\partial\theta}\right) \overset{\text{a.s.}}{\to} \mathrm{E}\left(\varepsilon_t\frac{\partial g(\theta, Y_{t+r})}{\partial\theta}\right) = \mathcal{J}_r,$$

where we use that ε_{t+r} and $\frac{\partial g(\theta,Y_t)}{\partial\theta}$ are independent. Subsituting this into (8.25) gives

$$\begin{aligned}\sqrt{n}\widehat{c}_n(r) &= \sqrt{n}\widetilde{c}_n(r) - \sqrt{n}(\widehat{\theta}-\theta)\mathcal{J}_r + o_p(1) \\ &= \sqrt{n}\widetilde{c}_n(r) - \mathcal{I}^{-1}\mathcal{J}_r\underbrace{\frac{1}{\sqrt{n}}\sum_{t=1}^{n-r}\frac{\partial g(Y_t,\theta)}{\partial\theta}\varepsilon_t}_{=-\frac{\sqrt{n}}{2}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}} + o_p(1).\end{aligned}$$

Asymptotic normality of $\sqrt{n}\widehat{c}_n(r)$ can be shown by showing asymptotic normality of the bivariate vector $\sqrt{n}(\widetilde{c}_n(r), \frac{\partial\mathcal{L}_n(\theta)}{\partial\theta})$. Therefore all that remains is to obtain the asymptotic variance of the above (which will give the desired result);

$$\begin{aligned}&\mathrm{var}\left[\sqrt{n}\widetilde{c}_n(r) + \frac{\sqrt{n}}{2}\mathcal{I}^{-1}\mathcal{J}_r\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\right] \\ &\underbrace{\mathrm{var}\left(\sqrt{n}\widetilde{c}_n(r)\right)}_{=1} + 2\mathcal{I}^{-1}\mathcal{J}_r\mathrm{cov}\left(\sqrt{n}\widetilde{c}_n(r), \frac{\sqrt{n}}{2}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\right) + \mathcal{I}^{-2}\mathcal{J}_r^2\mathrm{var}\left(\frac{\sqrt{n}}{2}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\right) \quad (8.26)\end{aligned}$$

We evaluate the two covariance above;

$$\begin{aligned}&\mathrm{cov}\left(\sqrt{n}\widetilde{c}_n(r), -\frac{\sqrt{n}}{2}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\right) = \frac{1}{n}\sum_{t_1,t_2=1}^{n-r}\left[\mathrm{cov}\left\{\varepsilon_{t_1}\varepsilon_{t_1+r}, \varepsilon_{t_2}\frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right\}\right] \\ = &\frac{1}{n}\sum_{t_1,t_2=1}^{n-r}\left[\mathrm{cov}\left\{\varepsilon_{t_1},\varepsilon_{t_2}\right\}\mathrm{cov}\left\{\varepsilon_{t_1+r}, \frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right\} + \mathrm{cov}\left\{\varepsilon_{t_1+r},\varepsilon_{t_2}\right\}\mathrm{cov}\left\{\varepsilon_{t_1}, \frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right\}\right. \\ &\left. + \mathrm{cum}\left\{\varepsilon_{t_1},\varepsilon_{t_1+r},\varepsilon_{t_2}, \frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right\}\right] = \sigma^2\mathrm{E}\left[\varepsilon_t\frac{\partial g(Y_{t+r},\theta)}{\partial\theta}\right] = \sigma^2\mathcal{J}_r.\end{aligned}$$

Similarly we have

$$\mathrm{var}\left(\frac{\sqrt{n}}{2}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\right) = \frac{1}{n}\sum_{t_1,t_2=1}^{n}\mathrm{cov}\left(\varepsilon_{t_1}\frac{\partial g(Y_{t_1},\theta)}{\partial\theta}, \varepsilon_{t_2}\frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right) = \sigma^2\mathrm{E}\left(\frac{\partial g(Y_{t_1},\theta)}{\partial\theta}\right)^2 = \sigma^2\mathcal{I}.$$

Substituting the above into (8.26) gives the asymptotic variance of $\sqrt{n}\widehat{c}(r)$ to be

$$1 - \sigma^2 \mathcal{J}_r \mathcal{I}^{-1} \mathcal{J}_r.$$

Thus we obtain the required result

$$\sqrt{n}\widehat{\rho}(r) = \mathcal{N}\left(0, 1 - \frac{\sigma^2}{c(0)} \mathcal{J}_r \mathcal{I}^{-1} \mathcal{J}_r\right).$$

8.7 Long range dependence (long memory) versus changes in the mean

A process is said to have long range dependence if the autocovariances are not absolutely summable, i.e. $\sum_k |c(k)| = \infty$. A nice historical background on long memory is given in this paper.

From a practical point of view data is said to exhibit long range dependence if the autocovariances do not decay very fast to zero as the lag increases. Returning to the Yahoo data considered in Section 13.1.1 we recall that the ACF plot of the absolute log differences, given again in Figure 8.4 appears to exhibit this type of behaviour. However, it has been argued by several authors that the 'appearance of long memory' is really because of a time-dependent mean has not been corrected for. Could this be the reason we see the 'memory' in the log differences?

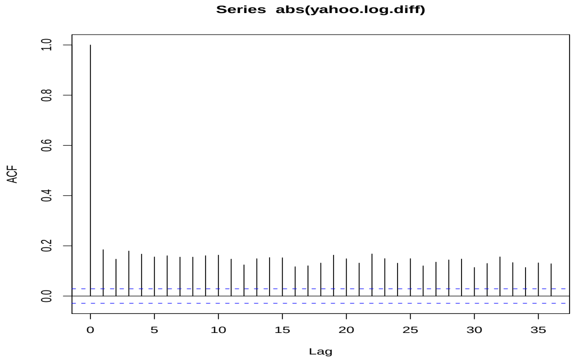


Figure 8.4: ACF plot of the absolute of the log differences.

We now demonstrate that one must be careful when diagnosing long range dependence, because a slow/none decay of the autocovariance could also imply a time-dependent mean that has not been corrected for. This was shown in Bhattacharya et al. (1983), and applied to econometric data in

Mikosch and Stărică (2000) and Mikosch and Stărică (2003). A test for distinguishing between long range dependence and change points is proposed in Berkes et al. (2006).

Suppose that Y_t satisfies

$$Y_t = \mu_t + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid random variables and the mean μ_t depends on t. We observe $\{Y_t\}$ but do not know the mean is changing. We want to evaluate the autocovariance function, hence estimate the autocovariance at lag k using

$$\hat{c}_n(k) = \frac{1}{n}\sum_{t=1}^{n-|k|}(Y_t - \bar{Y}_n)(Y_{t+|k|} - \bar{Y}_n).$$

Observe that $\bar{Y}_n$ is not really estimating the mean but the average mean! If we plotted the empirical ACF $\{\hat{c}_n(k)\}$ we would see that the covariances do not decay with time. However the true ACF would be zero and at all lags but zero. The reason the empirical ACF does not decay to zero is because we have not corrected for the time dependent mean. Indeed it can be shown that

$$\begin{aligned}
\hat{c}_n(k) &= \frac{1}{n}\sum_{t=1}^{n-|k|}(Y_t - \mu_t + \mu_t - \bar{Y}_n)(Y_{t+|k|} - \mu_{t+k} + \mu_{t+k} - \bar{Y}_n) \\
&\approx \frac{1}{n}\sum_{t=1}^{n-|k|}(Y_t - \mu_t)(Y_{t+|k|} - \mu_{t+k}) + \frac{1}{n}\sum_{t=1}^{n-|k|}(\mu_t - \bar{Y}_n)(\mu_{t+k} - \bar{Y}_n) \\
&\approx \underbrace{c(k)}_{\text{true autocovariance=0}} + \underbrace{\frac{1}{n}\sum_{t=1}^{n-|k|}(\mu_t - \bar{Y}_n)(\mu_{t+k} - \bar{Y}_n)}_{\text{additional term due to time-dependent mean}}
\end{aligned}$$

Expanding the second term and assuming that $k << n$ and $\mu_t \approx \mu(t/n)$ (and is thus smooth) we have

$$
\begin{aligned}
& \frac{1}{n}\sum_{t=1}^{n-|k|}(\mu_t-\bar{Y}_n)(\mu_{t+k}-\bar{Y}_n) \\
\approx\; & \frac{1}{n}\sum_{t=1}^{n}\mu_t^2-\left(\frac{1}{n}\sum_{t=1}^{n}\mu_t\right)^2+o_p(1) \\
=\; & \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\mu_t^2-\left(\frac{1}{n}\sum_{t=1}^{n}\mu_t\right)^2+o_p(1) \\
=\; & \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\mu_t\left(\mu_t-\mu_s\right)=\frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\mu_t-\mu_s\right)^2+\underbrace{\frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\mu_s\left(\mu_t-\mu_s\right)}_{=-\frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\mu_t(\mu_t-\mu_s)} \\
=\; & \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\mu_t-\mu_s\right)^2+\frac{1}{2n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\mu_s\left(\mu_t-\mu_s\right)-\frac{1}{2n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\mu_t\left(\mu_t-\mu_s\right) \\
=\; & \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\mu_t-\mu_s\right)^2+\frac{1}{2n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\mu_s-\mu_t\right)\left(\mu_t-\mu_s\right)=\frac{1}{2n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\mu_t-\mu_s\right)^2.
\end{aligned}
$$

Therefore

$$
\frac{1}{n}\sum_{t=1}^{n-|k|}(\mu_t-\bar{Y}_n)(\mu_{t+k}-\bar{Y}_n)\approx\frac{1}{2n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\mu_t-\mu_s\right)^2.
$$

Thus we observe that the sample covariances are positive and don't tend to zero for large lags. This gives the false impression of long memory.

It should be noted if you study a realisation of a time series with a large amount of dependence, it is unclear whether what you see is actually a stochastic time series or an underlying trend. This makes disentangling a trend from data with a large amount of correlation extremely difficult.

Chapter 9

Parameter estimation

Prerequisites

- The Gaussian likelihood.

Objectives

- To be able to derive the Yule-Walker and least squares estimator of the AR parameters.
- To understand what the quasi-Gaussian likelihood for the estimation of ARMA models is, and how the Durbin-Levinson algorithm is useful in obtaining this likelihood (in practice). Also how we can approximate it by using approximations of the predictions.
- Understand that there exists alternative methods for estimating the ARMA parameters, which exploit the fact that the ARMA can be written as an AR(∞).

We will consider various methods for estimating the parameters in a stationary time series. We first consider estimation parameters of an AR and ARMA process. It is worth noting that we will look at maximum likelihood estimators for the AR and ARMA parameters. The maximum likelihood will be constructed as if the observations were Gaussian. However, these estimators 'work' both when the process is Gaussian is also non-Gaussian. In the non-Gaussian case, the likelihood simply acts as a contrast function (and is commonly called the quasi-likelihood). In time series, often the distribution of the random variables is unknown and the notion of 'likelihood' has little meaning. Instead we seek methods that give good estimators of the parameters, meaning that they are consistent and as close to efficiency as possible without placing too many assumption on the distribution. We need to 'free' ourselves from the notion of likelihood acting as a likelihood (and attaining the Crámer-Rao lower bound).

9.1 Estimation for Autoregressive models

Let us suppose that $\{X_t\}$ is a zero mean stationary time series which satisfies the AR(p) representation

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t,$$

where $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{var}(\varepsilon_t) = \sigma^2$ and the roots of the characteristic polynomial $1 - \sum_{j=1}^{p} \phi_j z^j$ lie outside the unit circle. We will assume that the AR(p) is **causal** (the techniques discussed in this section cannot consistently estimate the parameters in the case that the process is non-causal, they will only consistently estimate the corresponding causal model). If you use the `ar` function in R to estimate the parameters, you will see that there are several different estimation methods that one can use to estimate $\{\phi_j\}$. These include, the Yule-Walker estimator, Least squares estimator, the Gaussian likelihood estimator and the Burg algorithm. Our aim in this section is to motivate and describe these several different estimation methods.

All these methods are based on their correlation structure. Thus they are only designed to estimate stationary, causal time series. For example, if we fit the AR(1) model $X_t = \phi X_{t-1} + \varepsilon_t$. The methods below cannot consistently estimate non-casual parameters (when $|\phi| > 1$). However, depending the method used, the estimator may be non-causal. For example, the classical least squares can yield estimators where $|\phi| > 1$. This does not mean the true model is non-causal, it simply means the minimum of the least criterion lies outside the parameter space $(-1, 1)$. Similarly, unless the parameter space of the MLE is constrained to only search for maximums inside $[1, 1]$ it can be give a maximum outside the natural parameter space. For the AR(1) estimator constraining the parameter space is quite simple. However, for higher order autoregressive models. Constraining the parameter space can be quite difficult.

On the other hand, both the Yule-Walker estimator and Burg's algorithm will always yield a causal estimator for any AR(p) model. There is no need to constrain the parameter space.

9.1.1 The Yule-Walker estimator

The Yule-Walker estimator is based on the Yule-Walker equations derived in (6.8) (Section 6.1.4).

We recall that the Yule-Walker equation state that if an AR process is causal, then for $i > 0$ we have

$$\mathrm{E}(X_t X_{t-i}) = \sum_{j=1}^{p} \phi_j \mathrm{E}(X_{t-j} X_{t-i}), \Rightarrow c(i) = \sum_{j=1}^{p} \phi_j c(i-j). \tag{9.1}$$

Putting the cases $1 \leq i \leq p$ together we can write the above as

$$\underline{r}_p = \Sigma_p \underline{\phi}_p, \tag{9.2}$$

where $(\Sigma_p)_{i,j} = c(i-j)$, $(\underline{r}_p)_i = c(i)$ and $\underline{\phi}_p' = (\phi_1, \ldots, \phi_p)$. Thus the autoregressive parameters solve these equations. It is important to observe that $\underline{\phi}_p = (\phi_1, \ldots, \phi_p)$ minimise the mean squared error

$$\mathrm{E}[X_{t+1} - \sum_{j=1}^{p} \phi_j X_{t+1-j}]^2,$$

(see Section 5.5).

The Yule-Walker equations inspire the method of moments estimator called the Yule-Walker estimator. We use (9.2) as the basis of the estimator. It is clear that $\hat{\underline{r}}_p$ and $\hat{\Sigma}_p$ are estimators of $\underline{r}_p$ and Σ_p where $(\hat{\Sigma}_p)_{i,j} = \hat{c}_n(i-j)$ and $(\hat{\underline{r}}_p)_i = \hat{c}_n(i)$. Therefore we can use

$$\hat{\underline{\phi}}_p = \hat{\Sigma}_p^{-1} \hat{\underline{r}}_p, \tag{9.3}$$

as an estimator of the AR parameters $\underline{\phi}_p' = (\phi_1, \ldots, \phi_p)$. We observe that if p is large this involves inverting a large matrix. However, we can use the Durbin-Levinson algorithm to calculate $\hat{\underline{\phi}}_p$ by recursively fitting lower order AR processes to the observations and increasing the order. This way an explicit inversion can be avoided. We detail how the Durbin-Levinson algorithm can be used to estimate the AR parameters below.

Step 1 Set $\hat{\phi}_{1,1} = \hat{c}_n(1)/\hat{c}_n(0)$ and $\hat{r}_n(2) = \hat{c}_n(0) - \hat{\phi}_{1,1}\hat{c}_n(1)$.

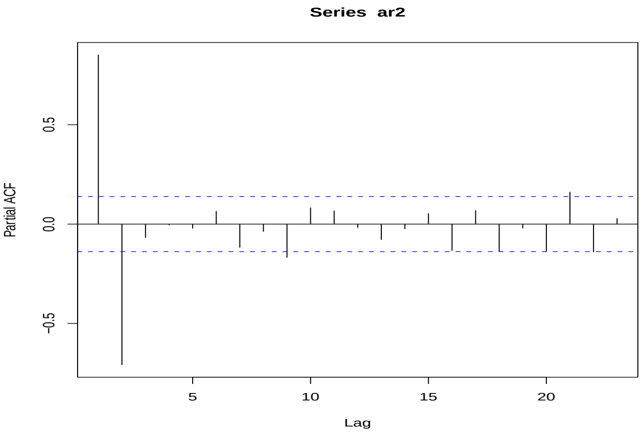


Figure 9.1: Top: The sample partial autocorrelation plot of the AR(2) process $X_t = 1.5X_{t-1} + 0.75X_{t-2} + \varepsilon_t$ with error bars $n = 200$.

Step 2 For $2 \leq t \leq p$, we define the recursion

$$\begin{aligned} \hat{\phi}_{t,t} &= \frac{\hat{c}_n(t) - \sum_{j=1}^{t-1} \hat{\phi}_{t-1,j}\hat{c}_n(t-j)}{\hat{r}_n(t)} \\ \hat{\phi}_{t,j} &= \hat{\phi}_{t-1,j} - \hat{\phi}_{t,t}\hat{\phi}_{t-1,t-j} \qquad 1 \leq j \leq t-1, \\ \text{and } \hat{r}_n(t+1) &= \hat{r}_n(t)(1-\hat{\phi}_{t,t}^2). \end{aligned}$$

Step 3 We recall from (7.11) that $\phi_{t,t}$ is the partial correlation between X_{t+1} and X_1, therefore $\hat{\phi}_{tt}$ are estimators of the partial correlation between X_{t+1} and X_1.

As mentioned in Step 3, the Yule-Walker estimators have the useful property that the partial correlations can easily be evaluated within the procedure. This is useful when trying to determine the order of the model to fit to the data. In Figure 9.1 we give the partial correlation plot corresponding to Figure 8.1. Notice that only the first two terms are outside the error bars. This rightly suggests the time series comes from an autoregressive process of order two.

In previous chapters it was frequently alluded to that the autocovariance is "blind" to non-causality and that any estimator based on estimating the covariance will always be estimating the causal solution. In Lemma 9.1.1 we show that the Yule-Walker estimator has the property that the parameter estimates $\{\hat{\phi}_j; j = 1, \ldots, p\}$ correspond to a causal AR(p), in other words, the roots corresponding to $\hat{\phi}(z) = 1 - \sum_{j=1}^{p} \widehat{\phi}_j z^j$ lie outside the unit circle. A non-causal solution cannot arise. The proof hinges on the fact that the Yule-Walker estimator is based on the sample autocovariances

$\{\widehat{c}_n(r)\}$ which are a positive semi-definite sequence (see Lemma 8.2.1).

Remark 9.1.1 (Fitting an AR(1) using the Yule-Walker) *We generalize this idea to general AR(p) models below. However, it is straightforward to show that the Yule-Walker estimator of the AR(1) parameter will always be less than or equal to one. We recall that*

$$\widehat{\phi}_{YW} = \frac{\sum_{t=1}^{n-1} X_t X_{t+1}}{\sum_{t=1}^{n} X_t^2}.$$

By using Cauchy-Schwarz we have

$$\begin{aligned} |\widehat{\phi}_{YW}| &\leq \frac{\sum_{t=1}^{n-1} |X_t X_{t+1}|}{\sum_{t=1}^{n} X_t^2} \leq \frac{[\sum_{t=1}^{n-1} X_t^2]^{1/2}[\sum_{t=1}^{n-1} X_{t+1}^2]^{1/2}}{\sum_{t=1}^{n} X_t^2} \\ &\leq \frac{[\sum_{t=1}^{n} X_t^2]^{1/2}[\sum_{t=0}^{n-1} X_{t+1}^2]^{1/2}}{\sum_{t=1}^{n} X_t^2} = 1. \end{aligned}$$

We use a similar idea below, but the proof hinges on the fact that the sample covariances forms a positive semi-definite sequence.

An alternative proof using that $\{\widehat{c}_n(r)\}$ is the ACF of a stationary time series $\{Z_t\}$. Then

$$\widehat{\phi}_{YW} = \frac{\widehat{c}_n(1)}{\widehat{c}_n(0)} = \frac{\mathrm{cov}(Z_t, Z_{t+1})}{\mathrm{var}(Z_t)} = \frac{\mathrm{cov}(Z_t, Z_{t+1})}{\sqrt{\mathrm{var}(Z_t)\mathrm{var}(Z_{t+1})}},$$

which is a correlation and thus lies between $[-1,1]$.

Lemma 9.1.1 *Let us suppose $\underline{Z}_{p+1} = (Z_1, \ldots, Z_{p+1})$ is a zero mean random vector, where* $\mathrm{var}[\underline{Z}]_{p+1} = (\Sigma_{p+1})_{i,j} = c_n(i-j)$ *(which is* **Toeplitz***). Let $Z_{p+1|p}$ be the best linear predictor of Z_{p+1} given $Z_p, \ldots, Z_1$, where $\underline{\phi}_p = (\phi_1, \ldots, \phi_p) = \Sigma_p^{-1}\underline{r}_p$ are the coefficients corresponding to the best linear predictor. Then the roots of the corresponding characteristic polynomial $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j$ lie outside the unit circle.*

PROOF. The proof is based on the following facts:

(i) Any sequence $\{\phi_j\}_{j=1}^{p}$ has the following reparameterisation. There exists parameters $\{a_j\}_{j=1}^{p}$ and λ such that $a_1 = 1$, for $2 \leq j \leq p-2$, $a_j - \lambda a_{j-1} = \phi_j$ and $\lambda a_p = \phi_p$. Using $\{a_j\}_{j=1}^{p}$ and λ, for rewrite the linear combination $\{Z_j\}_{j=1}^{p+1}$ as

$$Z_{p+1} - \sum_{j=1}^{p} \phi_j Z_{p+1-j} = \sum_{j=1}^{p} a_j Z_{p+1-j} - \lambda \sum_{j=1}^{p} a_j Z_{p-j}.$$

(ii) If $\underline{\phi}_p = (\phi_1, \ldots, \phi_p)' = \Sigma_p^{-1}\underline{r}_p$, then $\underline{\phi}_p$ minimises the mean square error i.e. for any $\{b_j\}_{j=1}^p$

$$\mathrm{E}_{\Sigma_{p+1}}\left(Z_{p+1} - \sum_{j=1}^p \phi_j Z_{p+1-j}\right)^2 \leq \mathrm{E}_{\Sigma_{p+1}}\left(Z_{p+1} - \sum_{j=1}^p b_j Z_{p+1-j}\right)^2 \tag{9.4}$$

where $\Sigma_{p+1} = \mathrm{var}[\underline{Z}_{p+1}]$ and $\underline{Z}_{p+1} = (Z_{p+1}, \ldots, Z_1)$.

We use these facts to prove the result. Our objective is to show that the roots of $\phi(B) = 1 - \sum_{j=1}^p \phi_j B^j$ lie outside the unit circle. Using (i) we factorize $\phi(B) = (1-\lambda B)a(B)$ where $a(B) = \sum_{j=1}^p a_j B^j$. Suppose by contraction $|\lambda| > 1$ (thus at least one root of $\phi(B)$ lies inside the unit circle). We will show if this were true, then by the Toeplitz nature of Σ_{p+1}, $\underline{\phi}_p = (\phi_1, \ldots, \phi_p)$ cannot be the best linear predictor.

Let

$$Y_{p+1} = \sum_{j=1}^p a_j B^j Z_{t+2} = \sum_{j=1}^p a_j Z_{p+2-j} \text{ and } Y_p = BY_{p+1} = B\sum_{j=1}^p a_j B^j Z_{t+2} = \sum_{j=1}^p a_j Z_{p+1-j}.$$

By (i) is clear that $Z_{p+1} - \sum_{j=1}^p \phi_j Z_{p+1-j} = Y_{p+1} - \lambda Y_p$. Furthermore, since $\{\phi_j\}$ minimises the mean squared error in (9.4), then λY_p must be the best linear predictor of Y_{p+1} given Y_p i.e. λ must minimise the mean squared error

$$\lambda = \arg\min_\beta \mathrm{E}_{\Sigma_{p+1}}\left(Y_{p+1} - \beta Y_p\right)^2,$$

that is $\lambda = \frac{\mathrm{E}[Y_{p+1}Y_p]}{\mathrm{E}[Y_p^2]}$. However, we now show that $|\frac{\mathrm{E}[Y_{p+1}Y_p]}{\mathrm{E}[Y_p^2]}| \leq 1$ which leads to a contradiction.

We recall that Y_{p+1} is a linear combination of a stationary sequence, thus BY_{p+1} has the same variance as Y_{p+1}. I.e. $\mathrm{var}(Y_{p+1}) = \mathrm{var}(Y_p)$. It you want to see the exact calculation, then

$$\begin{aligned}\mathrm{E}[Y_p^2] &= \mathrm{var}[Y_p] = \sum_{j_1,j_2=1}^p a_{j_1}a_{j_2}\mathrm{cov}[Y_{p+1-j_1}, Y_{p+1-j_2}] = \sum_{j_1,j_2=1}^p a_{j_1}a_{j_2}c(j_1-j_2) \\ &= \mathrm{var}[Y_{p+1}] = \mathrm{E}[Y_{p+1}^2].\end{aligned}$$

In other words, since Σ_{p+1} is a Toeplitz matrix, then $\mathrm{E}[Y_p^2] = \mathrm{E}[Y_{p+1}^2]$ and

$$\lambda = \frac{\mathrm{E}[Y_{p+1}Y_p]}{(\mathrm{E}[Y_p^2]\mathrm{E}[Y_{p+1}^2])^{1/2}}.$$

This means λ measures the correlation between Y_p and Y_{p+1} and must be less than or equal to one.

Thus leading to a contradiction.

Observe this proof only works when Σ_{p+1} is a Toeplitz matrix. If it is not we do not have $\mathrm{E}[Y_p^2] = \mathrm{E}[Y_{p+1}^2]$ and that λ can be intepretated as the correlation. $\square$

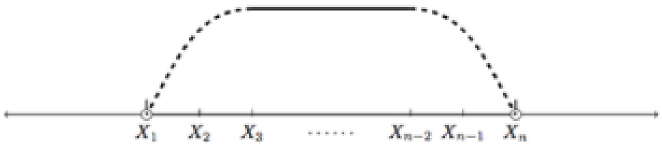
From the above result we can immediately see that the Yule-Walker estimators of the AR(p) coefficients yield a causal solution. Since the autocovariance estimators $\{\widehat{c}_n(r)\}$ form a positive semi-definite sequence, there exists a vector $\underline{Y}_p$ where $\mathrm{var}_{\widehat{\Sigma}_{p+1}}[\underline{Y}_{p+1}] = \widehat{\Sigma}_{p+1}$ with $(\widehat{\Sigma}_{p+1}) = \widehat{c}_n(i-j)$, thus by the above lemma we have that $\widehat{\Sigma}_p^{-1}\underline{\widehat{r}}_p$ are the coefficients of a Causal AR process.

Remark 9.1.2 (The bias of the Yule-Walker estimator) *The Yule-Walker tends to have larger bias than other other estimators when the sample size is small and the spectral density corresponding to the underlying time series is has a large pronounced peak (see Shaman and Stine (1988) and Ernst and Shaman (2019)). The large pronounced peak in the spectral density arises when the roots of the underlying characteristic polynomial lie close to the unit circle.*

9.1.2 The tapered Yule-Walker estimator

Substantial improvements to the Yule-Walker estimator can be obtained by tapering the original time series (tapering dates back to Tukey, but its application for AR(p) estimation was first proposed and proved in Dahlhaus (1988)).

Tapering is when the original data is downweighted towards the ends of the time series. This is done with a positive function $h : [0,1] \to \mathbb{R}$ that satisfies certain smoothness properties and is such that $h(0) = h(1) = 0$. And the tapered time series is $h(\frac{t}{n})X_t$. An illustration is given below:



In R, this can be done with the function `spec.taper(x,p=0.1)` where x is the time series, p is the proportion to be tapered). Replacing X_t with $h(t/n)X_t$ we define the tapered sample covariance as

$$\widehat{c}_{T,n}(r) = \frac{1}{\sum_{t=1}^{n} h(t/n)^2} \sum_{t=1}^{n-|r|} h\left(\frac{t}{n}\right) X_t h\left(\frac{t+r}{n}\right) X_{t+r}.$$

We now use $\{\widehat{c}_{T,n}(r)\}$ to define the Yule-Walker estimator for the AR(p) parameters.

9.1.3 The Gaussian likelihood

Our object here is to obtain the maximum likelihood estimator of the AR(p) parameters. We recall that the maximum likelihood estimator is the parameter which maximises the joint density of the observations. Since the log-likelihood often has a simpler form, we will focus on the log-likelihood. We note that the Gaussian MLE is constructed as if the observations $\{X_t\}$ were Gaussian, though it is not necessary that $\{X_t\}$ is Gaussian when doing the estimation. In the case that the innovations are not Gaussian, the estimator may be less efficient (may not obtain the Cramer-Rao lower bound) then the likelihood constructed as if the distribution were known.

Suppose we observe $\{X_t; t = 1, \ldots, n\}$ where X_t are observations from an AR(p) process. Let us suppose for the moment that the innovations of the AR process are Gaussian, this implies that $\underline{X}_n = (X_1, \ldots, X_n)$ is a n-dimension Gaussian random vector, with the corresponding log-likelihood

$$\mathcal{L}_n(\underline{a}) \quad = \quad -\log|\Sigma_n(\underline{a})| - \mathbf{X}_n'\Sigma_n(\underline{a})^{-1}\mathbf{X}_n, \tag{9.5}$$

where $\Sigma_n(\underline{a})$ the variance covariance matrix of $\mathbf{X}_n$ constructed as if $\mathbf{X}_n$ came from an AR process with parameters $\underline{a}$. Of course, in practice, the likelihood in the form given above is impossible to maximise. Therefore we need to rewrite the likelihood in a more tractable form.

We now derive a tractable form of the likelihood under the assumption that the innovations come from an arbitrary distribution. To construct the likelihood, we use the method of conditioning, to write the likelihood as the product of conditional likelihoods. In order to do this, we derive the conditional distribution of X_{t+1} given $X_{t-1}, \ldots, X_1$. We first note that the AR(p) process is p-Markovian (if it is causal), therefore if $t \geq p$ all the information about X_{t+1} is contained in the past p observations, therefore

$$\mathbb{P}(X_{t+1} \leq x|X_t, X_{t-1}, \ldots, X_1) = \mathbb{P}(X_{t+1} \leq x|X_t, X_{t-1}, \ldots, X_{t-p+1}), \tag{9.6}$$

by causality. Since the Markov property applies to the distribution function it also applies to the density

$$f(X_{t+1}|X_t, \ldots, X_1) = f(X_{t+1}|X_t, \ldots, X_{t-p+1}).$$

By using the (9.6) we have

$$\mathbb{P}(X_{t+1} \leq x|X_t, \ldots, X_1) = \mathbb{P}(X_{t+1} \leq x|X_t, \ldots, X_1) = \mathbb{P}_{\varepsilon}(\varepsilon \leq x - \sum_{j=1}^{p} a_j X_{t+1-j}), \tag{9.7}$$

where $\mathbb{P}_{\varepsilon}$ denotes the distribution of the innovation. Differentiating $\mathbb{P}_{\varepsilon}$ with respect to X_{t+1} gives

$$f(X_{t+1}|X_t, \ldots, X_{t-p+1}) = \frac{\partial \mathbb{P}_{\varepsilon}(\varepsilon \leq X_{t+1} - \sum_{j=1}^{p} a_j X_{t+1-j})}{\partial X_{t+1}} = f_{\varepsilon}\left(X_{t+1} - \sum_{j=1}^{p} a_j X_{t+1-j}\right). \tag{9.8}$$

Example 9.1.1 (AR(1)) *To understand why (9.6) is true consider the simple case that $p = 1$ ($AR(1)$ with $|\phi| < 1$). Studying the conditional probability gives*

$$\begin{aligned}\mathbb{P}(X_{t+1} \leq x_{t+1}|X_t = x_t, \ldots, X_1 = x_1) &= \mathbb{P}(\underbrace{\phi X_t + \varepsilon_t \leq x_{t+1}}_{\textit{all information contained in } X_t} |X_t = x_t, \ldots, X_1 = x_1) \\ &= \mathbb{P}_{\varepsilon}(\varepsilon_t \leq x_{t+1} - \phi x_t) = \mathbb{P}(X_{t+1} \leq x_{t+1}|X_t = x_t),\end{aligned}$$

where $\mathbb{P}_{\varepsilon}$ denotes the distribution function of the innovation ε.

Using (9.8) we can derive the joint density of $\{X_t\}_{t=1}^n$. By using conditioning we obtain

$$\begin{aligned}f(X_1, X_2, \ldots, X_n) &= f(X_1, \ldots, X_p) \prod_{t=p}^{n-1} f(X_{t+1}|X_t, \ldots, X_1) \quad \text{(by repeated conditioning)} \\ &= f(X_1, \ldots, X_p) \prod_{t=p}^{n-1} f(X_{t+1}|X_t, \ldots, X_{t-p+1}) \quad \text{(by the Markov property)} \\ &= f(X_1, \ldots, X_p) \prod_{t=p}^{n-1} f_{\varepsilon}(X_{t+1} - \sum_{j=1}^{p} a_j X_{t+1-j}) \quad \text{(by (9.8))}.\end{aligned}$$

Therefore the log likelihood is

$$\underbrace{\log f(X_1, X_2, \ldots, X_n)}_{\text{Full log-likelihood } \mathcal{L}_n(\underline{a};\underline{X}_n)} = \underbrace{\log f(X_1, \ldots, X_p)}_{\text{initial observations}} + \underbrace{\sum_{t=p}^{n-1} \log f_{\varepsilon}(X_{t+1} - \sum_{j=1}^{p} a_j X_{t+1-j})}_{\text{conditional log-likelihood}=L_n(\underline{a};\underline{X}_n)}.$$

In the case that the sample sizes are large $n >> p$, the contribution of initial observations $\log f(X_1, \ldots, X_p)$ is minimal and the conditional log-likelihood and full log-likelihood are asymptotically equivalent.

So far we have not specified the distribution of $\{\varepsilon_t\}_t$. From now on we shall assume that it is

Gaussian. Thus $\log f(X_1,\ldots,X_n;\phi)$ and $\log f(X_1,\ldots,X_p;\phi)$ are multivariate normal with mean zero (since we are assuming, for convenience, that the time series has zero mean) and variance $\Sigma_n(\phi)$ and $\Sigma_p(\phi)$ respectively, where by stationarity $\Sigma_n(\phi)$ and $\Sigma_p(\phi)$ are Toeplitz matrices. Based on this the (negative) log-likelihood is

$$\begin{aligned}
\mathcal{L}_n(\underline{a}) &= \log|\Sigma_n(\underline{a})| + \underline{X}_p'\Sigma_n(\underline{a})^{-1}\underline{X}_p \\
&= \log|\Sigma_p(\underline{a})| + \underline{X}_p'\Sigma_p(\underline{a})^{-1}\underline{X}_p + \underbrace{L_n(\underline{a};\underline{X})}_{\text{conditional likelihood}} . \qquad (9.9)
\end{aligned}$$

The maximum likelihood estimator is

$$\widehat{\underline{\phi}}_n = \arg\max_{\underline{a}\in\Theta}\mathcal{L}_n(\underline{a}). \qquad (9.10)$$

The parameters in the model are 'buried' within the covariance. By constraining the parameter space, we can ensure the estimator correspond to a causal AR process (but find suitable parameter space is not simple). Analytic expressions do exist for $\underline{X}_p'\Sigma_p(\underline{a})^{-1}\underline{X}_p$ and $\log|\Sigma_p(\underline{a})|$ but they are not so simple. This motivates the conditional likelihood described in the next section.

9.1.4 The conditional Gaussian likelihood and least squares

The conditonal likelihood focusses on the conditonal term of the Gaussian likelihood and is defined as

$$L_n(\underline{a};\underline{X}) = -(n-p)\log\sigma^2 - \frac{1}{\sigma^2}\sum_{t=p}^{n-1}\left(X_{t+1}-\sum_{j=1}^{p}a_jX_{t+1-j}\right)^2,$$

is straightforward to maximise. Since the maximum of the above with respect to $\{a_j\}$ does not depend on σ^2. The conditional likelihood estimator of $\{\phi_j\}$ is simply the least squares estimator

$$\begin{aligned}
\widetilde{\underline{\phi}}_p &= \arg\min\sum_{t=p}^{n-1}\left(X_{t+1}-\sum_{j=1}^{p}a_jX_{t+1-j}\right)^2 \\
&= \widetilde{\Sigma}_p^{-1}\widetilde{\underline{r}}_p,
\end{aligned}$$

where $(\widetilde{\Sigma}_p)_{i,j} = \frac{1}{n-p}\sum_{t=p+1}^{n}X_{t-i}X_{t-j}$ and $(\widetilde{\underline{r}}_n)_i = \frac{1}{n-p}\sum_{t=p+1}^{n}X_tX_{t-i}$.

Remark 9.1.3 (A comparison of the Yule-Walker and least squares estimators) *Comparing* *the least squares estimator* $\widetilde{\underline{\phi}}_p = \widetilde{\Sigma}_p^{-1}\widetilde{\underline{r}}_p$ *with the Yule-Walker estimator* $\widehat{\underline{\phi}}_p = \widehat{\Sigma}_p^{-1}\widehat{\underline{r}}_p$ *we see that they are very similar. The difference lies in* $\widetilde{\Sigma}_p$ *and* $\widehat{\Sigma}_p$ *(and the corresponding* $\widetilde{\underline{r}}_p$ *and* $\widehat{\underline{r}}_p$*). We see that* $\widehat{\Sigma}_p$ *is a Toeplitz matrix, defined entirely by the positive definite sequence* $\widehat{c}_n(r)$*. On the other hand,* $\widetilde{\Sigma}_p$ *is not a Toeplitz matrix, the estimator of* $c(r)$ *changes subtly at each row. This means that the proof given in Lemma 9.1.1 cannot be applied to the least squares estimator as it relies on the matrix* Σ_{p+1} *(which is a combination of* Σ_p *and* $\underline{r}_p$*) being Toeplitz (thus stationary). Thus the characteristic polynomial corresponding to the least squares estimator will not necessarily have roots which lie outside the unit circle.*

Example 9.1.2 (Toy Example) *To illustrate the difference between the Yule-Walker and least squares estimator (at least for example samples) consider the rather artifical example that the time series consists of two observations* X_1 *and* X_2 *(we will assume the mean is zero). We fit an* $AR(1)$ *model to the data, the least squares estimator of the* $AR(1)$ *parameter is*

$$\widehat{\phi}_{LS} = \frac{X_1X_2}{X_1^2}$$

whereas the Yule-Walker estimator of the $AR(1)$ *parameter is*

$$\widehat{\phi}_{YW} = \frac{X_1X_2}{X_1^2 + X_2^2}.$$

It is clear that $\widehat{\phi}_{LS} < 1$ *only if* $X_2 < X_1$*. On the other hand* $\widehat{\phi}_{YW} < 1$*. Indeed since* $(X_1 - X_2)^2 > 0$*, we see that* $\widehat{\phi}_{YW} \leq 1/2$*.*

Exercise 9.1 *(i) In* `R` *you can estimate the AR parameters using ordinary least squares (*`ar.ols`*), yule-walker (*`ar.yw`*) and (Gaussian) maximum likelihood (*`ar.mle`*).*

Simulate the causal $AR(2)$ *model* $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$ *using the routine* `arima.sim` *(which gives Gaussian realizations) and also innovations which from a t-distribution with 4df. Use the sample sizes* $n = 100$ *and* $n = 500$ *and compare the three methods through a simulation study.*

(ii) Use the ℓ_1*-norm defined as*

$$L_n(\phi) = \sum_{t=p+1}^{t} \left| X_t - \sum_{j=1}^{p} \phi_j X_{t-j} \right|,$$

with $\hat{\phi}_n = \arg\min L_n(\phi)$ to estimate the AR(p) parameters.

You may need to use a Quantile Regression package to minimise the ℓ_1 norm. I suggest using the package `quantreg` *and the function* `rq` *where we set $\tau = 0.5$ (the median).*

Note that so far we have only considered estimation of causal AR(p) models. Breidt et. al. (2001) propose a method for estimating parameters of a non-causal AR(p) process (see page 18).

9.1.5 Burg's algorithm

Burg's algorithm is an alternative method for estimating the AR(p) parameters. It is closely related to the least squares estimator but uses properties of second order stationarity in its construction. Like the Yule-Walker estimator it has the useful property that its estimates correspond to a causal characteristic function. Like the Yule-Walker estimator it can recursively estimate the AR(p) parameters by first fitting an AR(1) model and then recursively increasing the order of fit.

We start with fitting an AR(1) model to the data. Suppose that $\phi_{1,1}$ is the true best fitting AR(1) parameter, that is

$$X_t = P_{X_{t-1}}(X_t) + \varepsilon_{1,t} = \phi_{1,1}X_{t-1} + \varepsilon_{1,t}.$$

Then the least squares estimator is based on estimating the projection by using the $\phi_{1,1}$ that minimises

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

However, the same parameter $\phi_{1,1}$ minimises the projection of the future into the past

$$X_t = P_{X_{t+1}}(X_t) + \delta_{1,t} = \phi_{1,1}X_{t+1} + \delta_{1,t}.$$

Thus by the same argument as above, an estimator of $\phi_{1,1}$ is the parameter which minimises

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

We can combine these two least squares estimators to find the ϕ which minimises

$$\widehat{\phi}_{1,1} = \arg\min \left[\sum_{t=2}^{n} (X_t - \phi X_{t-1})^2 + \sum_{t=1}^{n-1} (X_t - \phi X_{t+1})^2 \right].$$

Differentiating the above wrt ϕ and solving gives the explicit expression

$$\begin{aligned} \widehat{\phi}_{1,1} &= \frac{\sum_{t=1}^{n-1} X_t X_{t+1} + \sum_{t=2}^{n} X_t X_{t-1}}{2\sum_{t=2}^{n-1} X_t^2 + X_1^2 + X_n^2} \\ &= \frac{2\sum_{t=1}^{n-1} X_t X_{t+1}}{2\sum_{t=2}^{n-1} X_t^2 + X_1^2 + X_n^2}. \end{aligned}$$

Unlike the least squares estimator $\widehat{\phi}_{1,1}$ is guaranteed to lie between $[-1, 1]$. Note that $\phi_{1,1}$ is the partial correlation at lag one, thus $\widehat{\phi}_{1,1}$ is an estimator of the partial correlation. In the next step we estimate the partial correlation at lag two. We use the projection argument described in Sections 5.1.4 and 7.5.1. That is

$$P_{X_{t-2},X_{t-1}}(X_t) = P_{X_{t-1}}(X_t) + \rho \left(X_{t-2} - P_{X_{t-1}}(X_{t-2}) \right)$$

and

$$\begin{aligned} X_t &= P_{X_{t-2},X_{t-1}}(X_t) + \varepsilon_{2,t} = P_{X_{t-1}}(X_t) + \rho \left(X_{t-2} - P_{X_{t-1}}(X_{t-2}) \right) + \varepsilon_{2,t} \\ &= \phi_{1,1} X_{t-1} + \rho \left(X_{t-2} - \phi_{1,1} X_{t-1} \right) + \varepsilon_{2,t}. \end{aligned}$$

Thus we replace $\phi_{1,1}$ in the above with $\widehat{\phi}_{1,1}$ and estimate ρ by minimising least squares criterion

$$\sum_{t=3}^{n} \left[X_t - \widehat{\phi}_{1,1} X_{t-1} - \rho \left(X_{t-2} - \widehat{\phi}_{1,1} X_{t-1} \right) \right].$$

However, just as in the estimation scheme of $\phi_{1,1}$ we can estimate ρ by predicting into the past

$$P_{X_{t+2},X_{t+1}}(X_t) = P_{X_{t+1}}(X_t) + \rho \left(X_{t+2} - P_{X_{t+1}}(X_{t+2}) \right)$$

to give

$$X_t = \phi_{1,1} X_{t+1} + \rho \left(X_{t+2} - \phi_{1,1} X_{t+1} \right) + \delta_{2,t}.$$

This leads to an alternative estimator of ρ that minimises

$$\sum_{t=1}^{n-2} \left[X_t - \widehat{\phi}_{1,1} X_{t+1} - \rho \left(X_{t+2} - \widehat{\phi}_{1,1} X_{t+1} \right) \right].$$

The Burg algorithm estimator of ρ minimises both the forward and backward predictor simultaneously

$$\widehat{\rho}_2 = \arg\min_{\rho} \left(\sum_{t=3}^{n} \left[X_t - \widehat{\phi}_{1,1} X_{t-1} - \rho \left(X_{t-2} - \widehat{\phi}_{1,1} X_{t-1} \right) \right] + \sum_{t=1}^{n-2} \left[X_t - \widehat{\phi}_{1,1} X_{t+1} - \rho \left(X_{t+2} - \widehat{\phi}_{1,1} X_{t+1} \right) \right] \right).$$

Differentiating the above wrt ρ and solving gives an explicit solution for $\widehat{\rho}_2$. Moreover we can show that $|\widehat{\rho}_2| \leq 1$. The estimators of the best fitting AR(2) parameters $(\phi_{1,2}, \phi_{2,2})$ are

$$\begin{aligned} \widehat{\phi}_{1,2} &= \left(\widehat{\phi}_{1,1} - \widehat{\rho}_2 \widehat{\phi}_{1,1} \right) \\ \text{and } \widehat{\phi}_{2,2} &= \widehat{\rho}_2. \end{aligned}$$

Using the same method we can obtain estimators for $\{\widehat{\phi}_{r,r}\}_r$ which can be used to construct the estimates of the best fitting AR(p) parameters $\{\widehat{\phi}_{j,p}\}_{j=1}^{p}$. It can be shown that the parameters $\{\widehat{\phi}_{j,p}\}_{j=1}^{p}$ correspond to a causal AR(p) model.

Proof that $0 \leq |\widehat{\phi}_{1,1}| \leq 1$ To prove the result we pair the terms in the estimator

$$\widehat{\phi}_{1,1} = \frac{2\left[X_1X_2 + X_2X_3 + \ldots + X_{n-1}X_n\right]}{(X_1^2 + X_2^2) + (X_2^2 + X_3^2) + \ldots + (X_{n-2}^2 + X_{n-1}^2) + (X_{n-1}^2 + X_n^2)}.$$

Each term in the numerator can be paired with the term in the denominator i.e. using that $(|X_t| - |X_{t+1}|)^2 \geq 0$ we have

$$2|X_tX_{t+1}| \leq X_t^2 + X_{t+1}^2 \qquad 1 \leq t \leq (n-1).$$

Thus the absolute of the numerator is smaller that the denominator and we have

$$|\widehat{\phi}_{1,1}| = \frac{2\left[|X_1X_2| + |X_2X_3| + \ldots + |X_{n-1}X_n|\right]}{(X_1^2 + X_2^2) + (X_2^2 + X_3^2) + \ldots + (X_{n-2}^2 + X_{n-1}^2) + (X_{n-1}^2 + X_n^2)} \leq 1.$$

This proves the claim. □