A course in Time Series Analysis

Suhasini Subba Rao

Email: suhasini.subbarao@stat.tamu.edu

January 17, 2021

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.
  - A whole bunch of articles.
- 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.

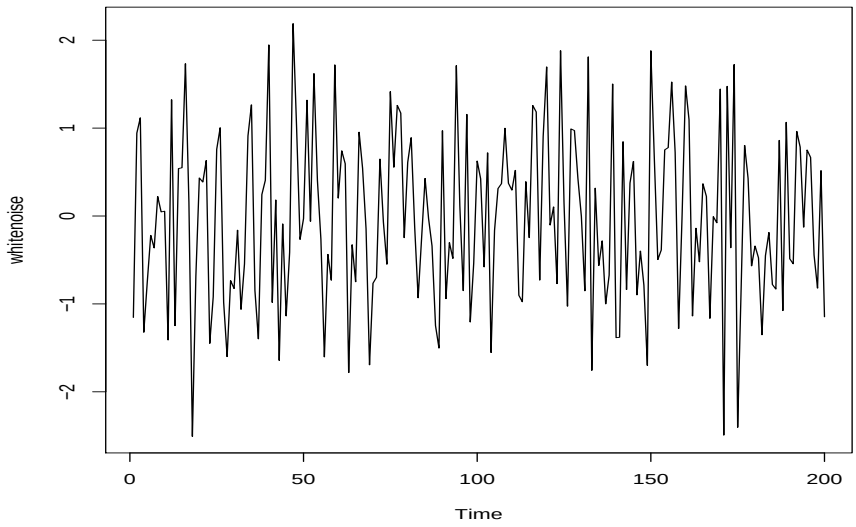


Figure 1.1: Plot of independent uncorrelated random variables

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.

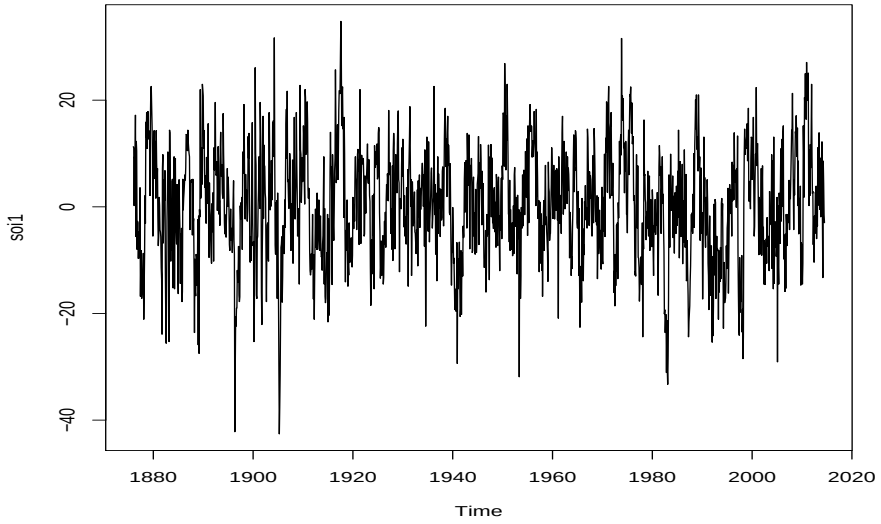


Figure 1.2: Plot of monthly Southern Oscillation Index, 1876-2014

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).

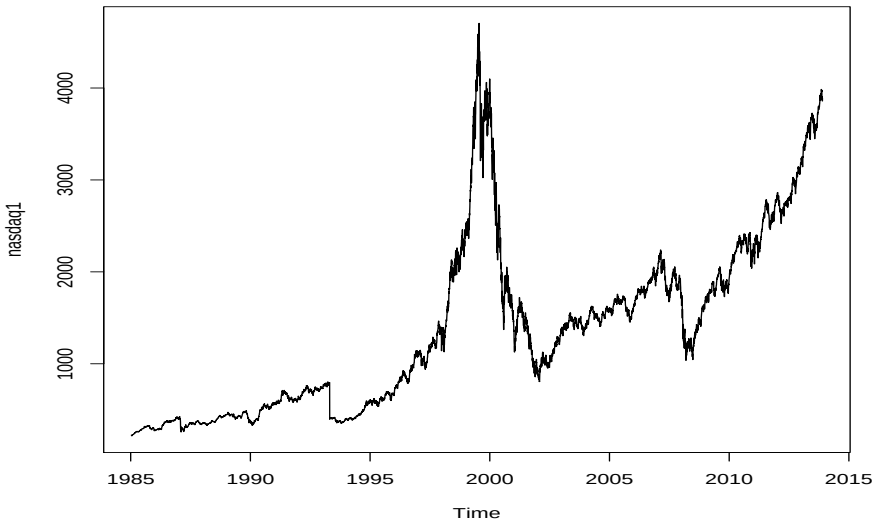


Figure 1.3: Plot of daily closing price of Nasdaq 1985-2014

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

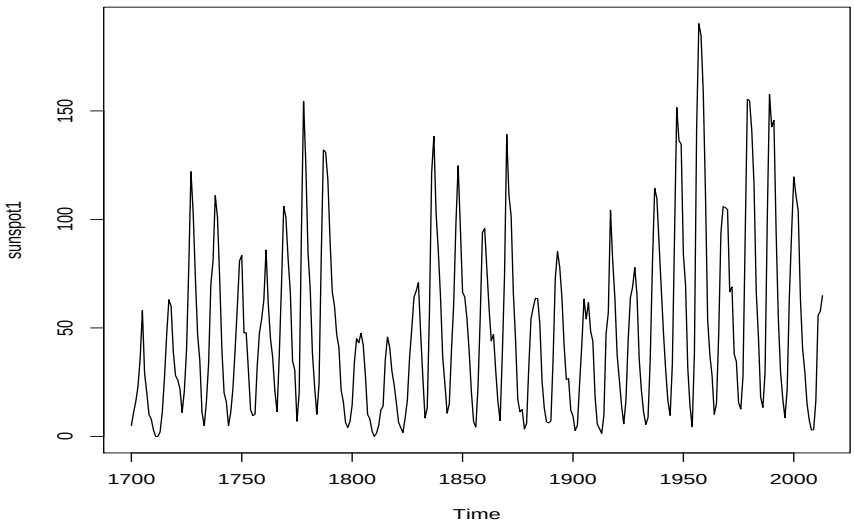


Figure 1.4: Plot of Sunspot numbers 1700-2013

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,
```

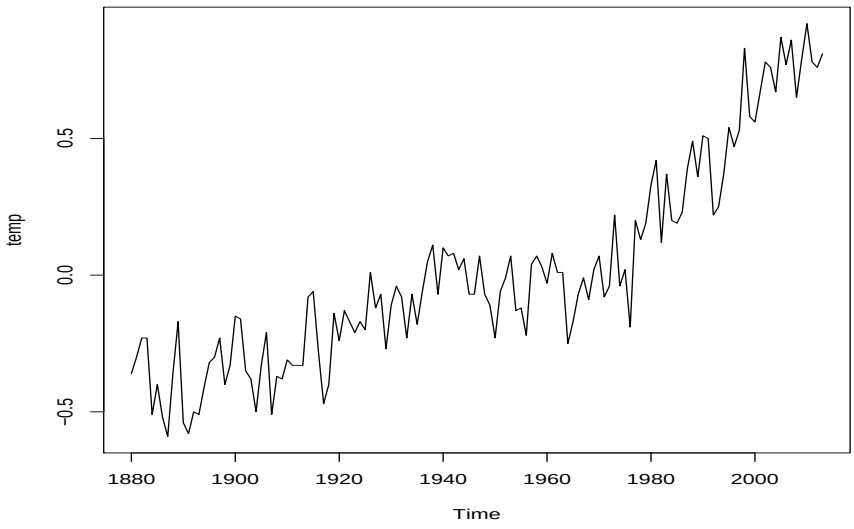


Figure 1.5: Plot of global, yearly average, temperature anomalies, 1880 - 2013

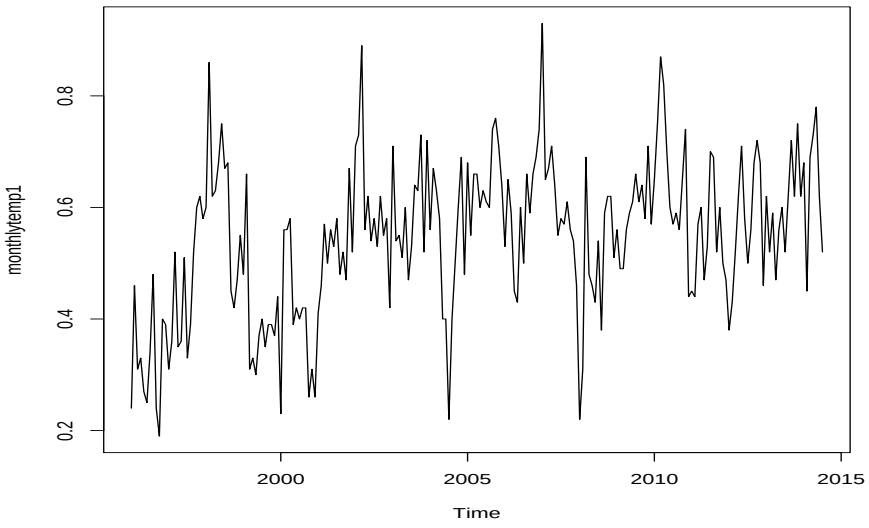


Figure 1.6: Plot of global, monthly average, temperatures January, 1996 - July, 2014.

```
# 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

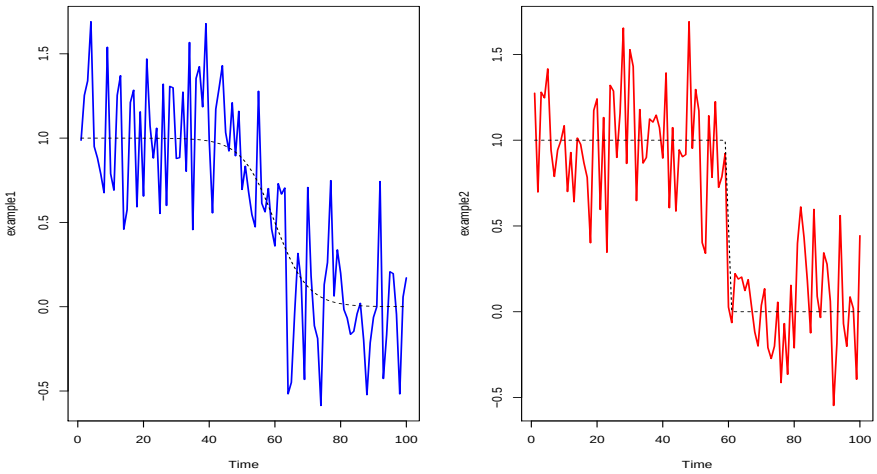


Figure 2.1: The function Y_t in (2.3) with iid noise with $\sigma = 0.3$. Dashed is the truth. Left: $\beta_0 = 0.2$ and $\beta_1 = 60$. Right: $\beta_0 = 5$ and $\beta_1 = 60$

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

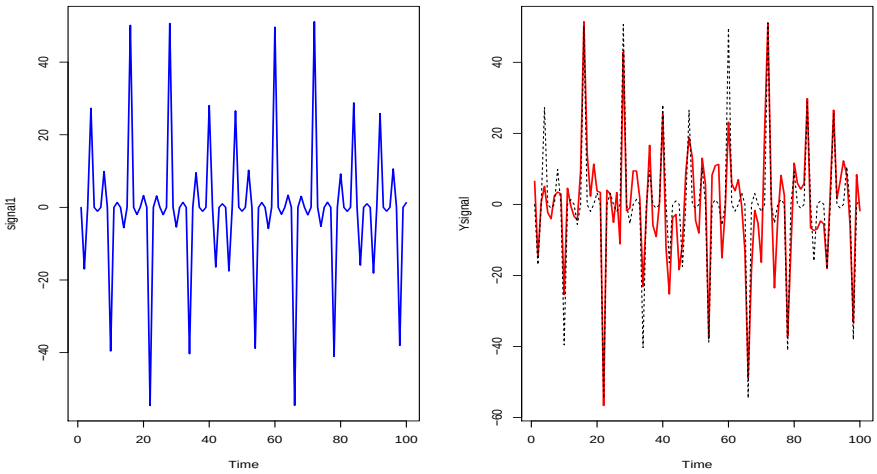


Figure 2.2: The Burst signal (equation (2.4)) $A = 1$, $\beta_0 = 2$, $\beta_1 = 1$ and $\theta = \pi/2$ with iid noise with $\sigma = 8$. Dashed is the truth. Left: True Signal. Right: True Signal with noise

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

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

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. $\mathrm{cov}[\varepsilon_i, X_i]$. 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* $Y_i =$

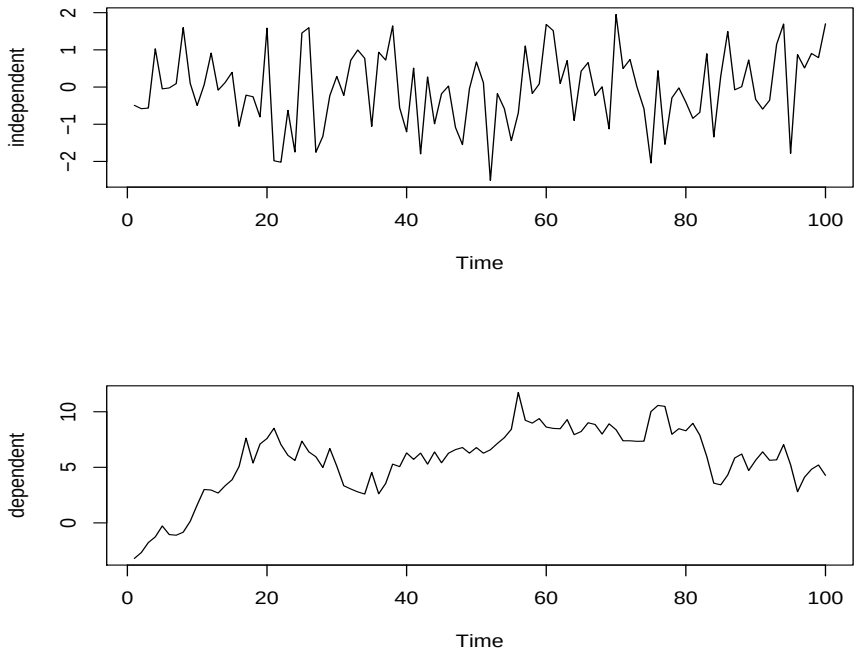


Figure 2.3: Top: realisations from iid random noise. Bottom: Realisation from dependent noise

$f(i/200) + \varepsilon_i$ *(*$n = 200$*). 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.

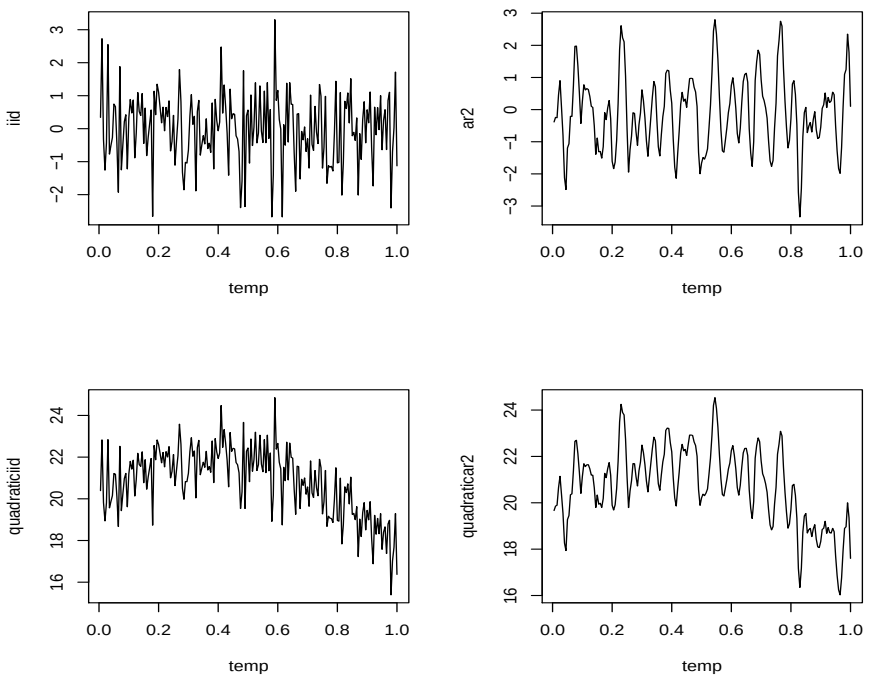


Figure 2.4: Top: realisations from iid random noise and dependent noise (left = iid and right = dependent). Bottom: Quadratic trend plus corresponding noise.

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

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

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

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

These are often called the sin and cosine transforms.

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

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

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 $\cos(\omega) = \mathrm{Re}\, e^{i\omega} = (e^{i\omega} + e^{-i\omega})/2$ and $\sin(\omega) = \mathrm{Im}\, e^{i\omega} = -i(e^{i\omega} - e^{-i\omega})/2$.

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 $d_P(t) =$

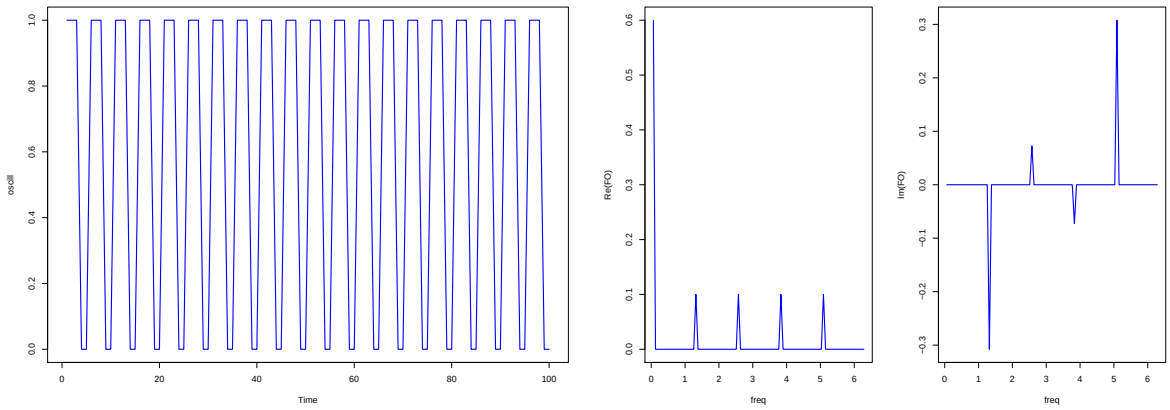


Figure 2.5: Left: Periodic function $d_5(s) = 1$ for $s = 1, 2$, $d_5(s) = 0$ for $s = 3, 4, 5$ (period 5), Right: The real and imaginary parts of its Fourier transform

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

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

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:

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

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

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

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

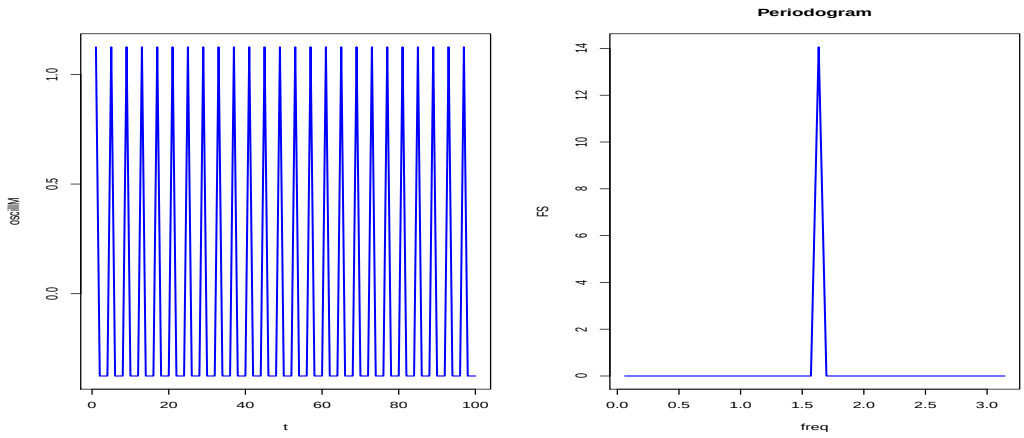


Figure 2.6: Left: Periodic function $d_4(s) = (1.125, -0.375, -0.375, -0.375)$ (period 4)

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.

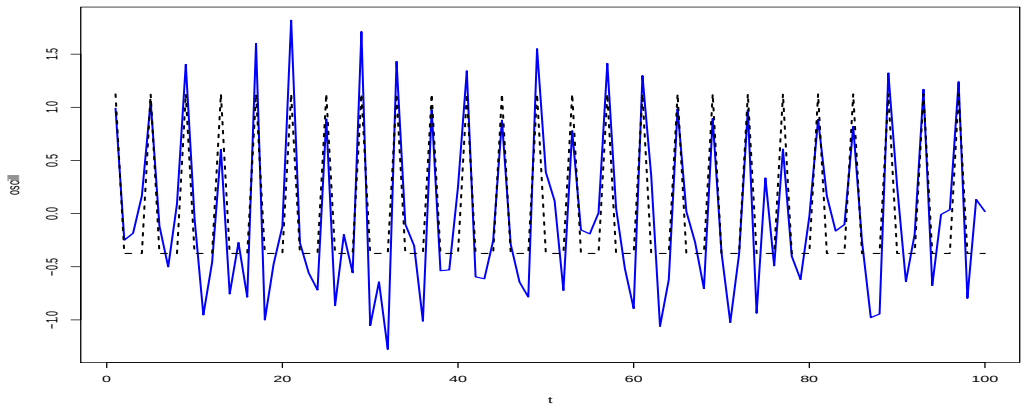


Figure 2.7: Periodic function $d_4(s) = (1.125, -0.375, -0.375, -0.375)$ (period 4) and signal with noise (blue line).

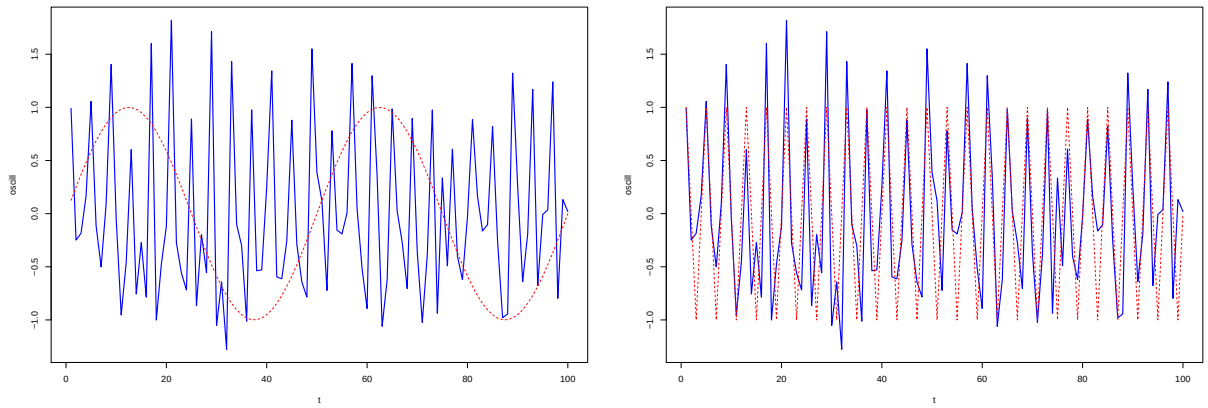


Figure 2.8: Left: Signal superimposed with $\sin(u)$. Right: Signal superimposed with $\sin(25u)$.

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

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

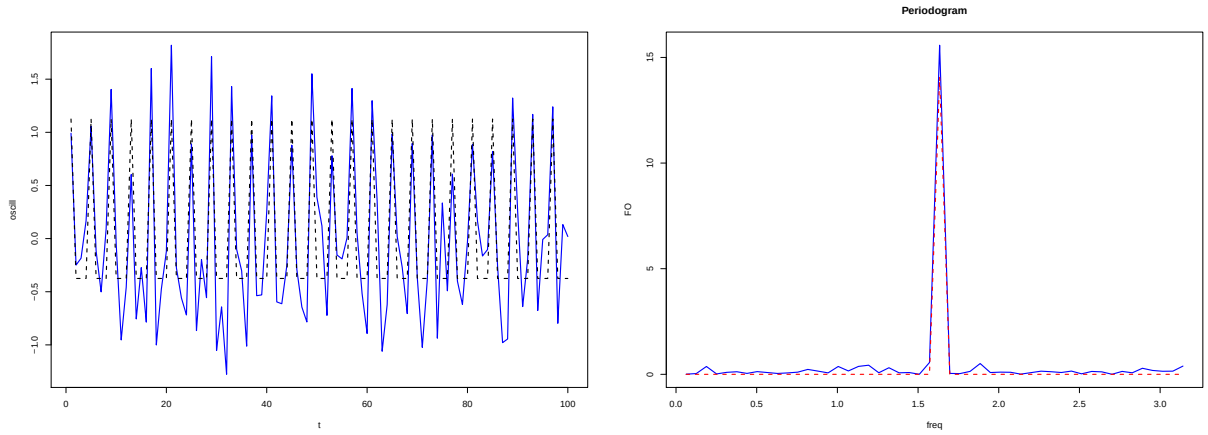


Figure 2.9: Left: Signal, Right: periodogram of signal (peridogram of periodic function in red)

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 $\text{var}[\varepsilon_t] = 0.5^2$. 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).

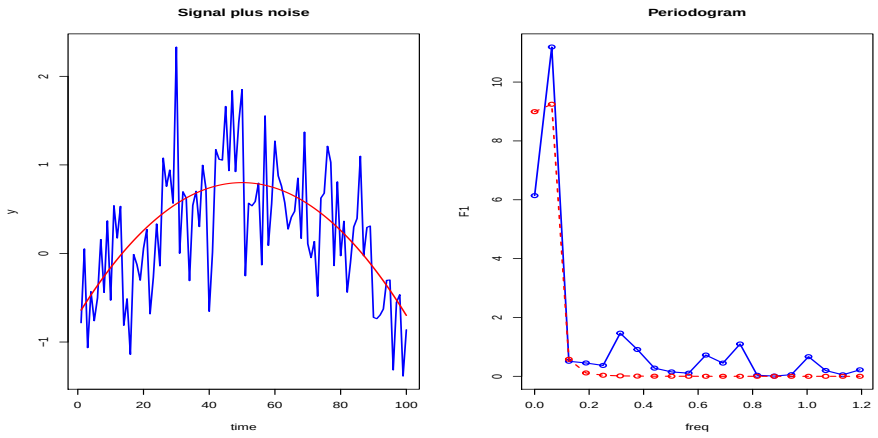


Figure 2.10: Left: Signal and noise (blue). The signal is in red. Right: Periodogram of signal plus noise (up to frequency $\pi/5$). Periodogram of signal is in red.

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

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

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*

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

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*).*

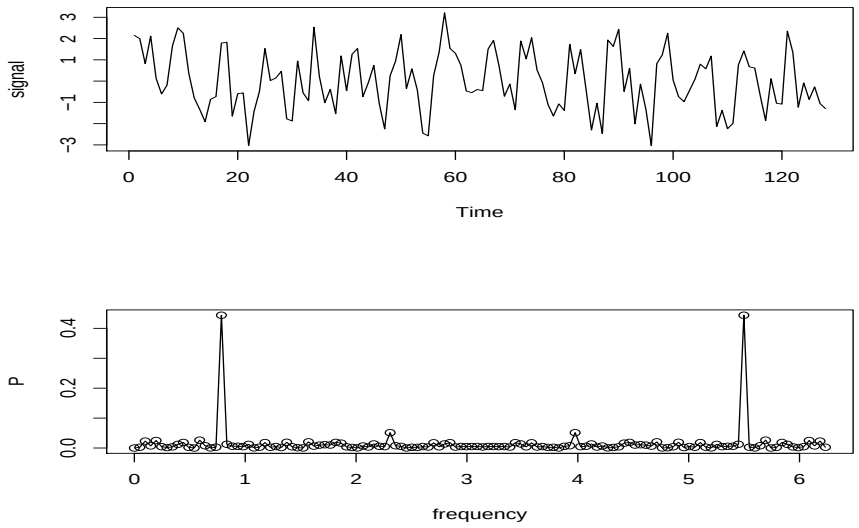


Figure 2.11: Left: Realisation of (2.18) plus iid noise, Right: Periodogram of signal plus iid noise.

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 <u>not</u> 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

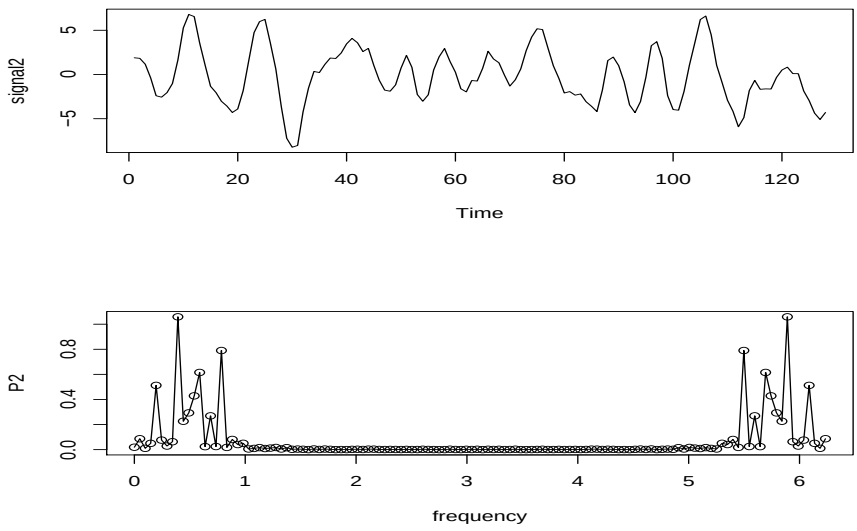


Figure 2.12: Top: Realisation of (2.18) plus correlated noise and $n = 128$, Bottom: Periodogram of signal plus correlated noise.

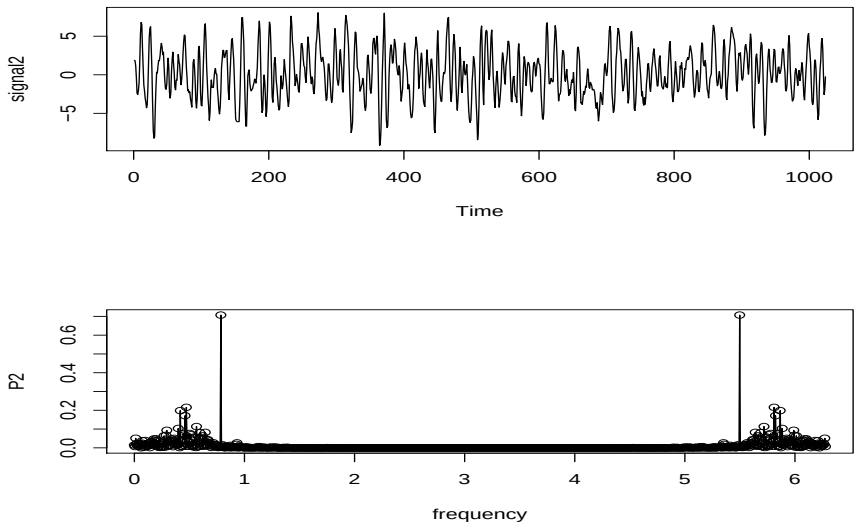


Figure 2.13: Top: Realisation of (2.18) plus correlated noise and $n = 1024$, Bottom: Periodogram of signal plus correlated noise.

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

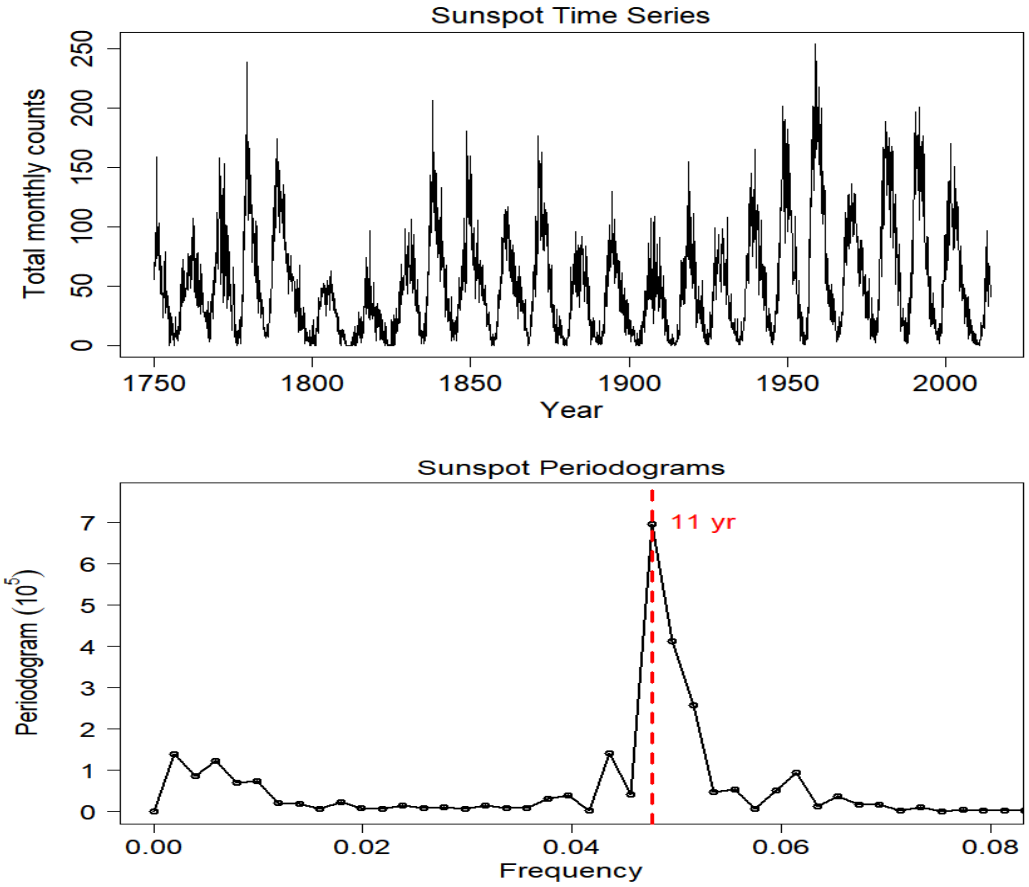


Figure 2.14: Sunspot data from Jan, 1749 to Dec, 2014. There is a peak at about 30 along the line which corresponds to $2\pi/P = 0.047$ and $P \approx 132$ months (11 years).

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

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

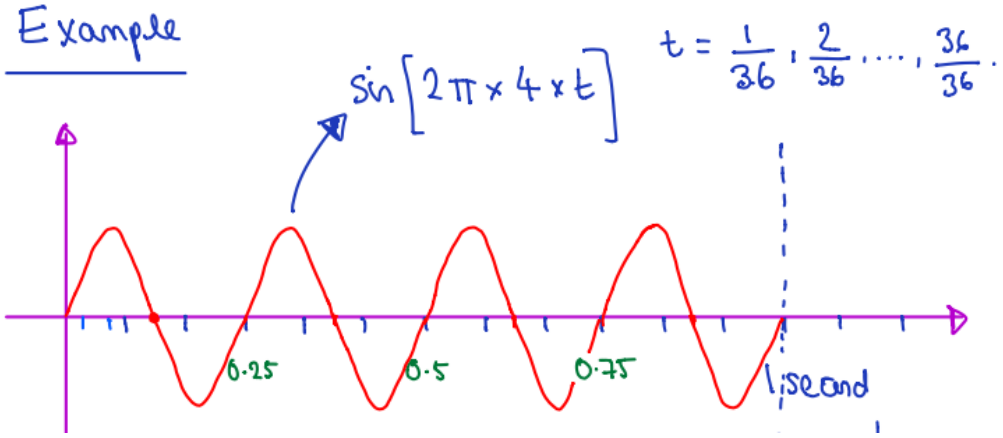
However, in the 1920's, Udny Yule, a statistician, and Gilbert Walker, a Meterologist (working in Pune, India) believed an alternative model could be used to explain the features seen in the periodogram. We consider their proposed approach in Section 4.3.5.

2.6 Data Analysis: EEG data

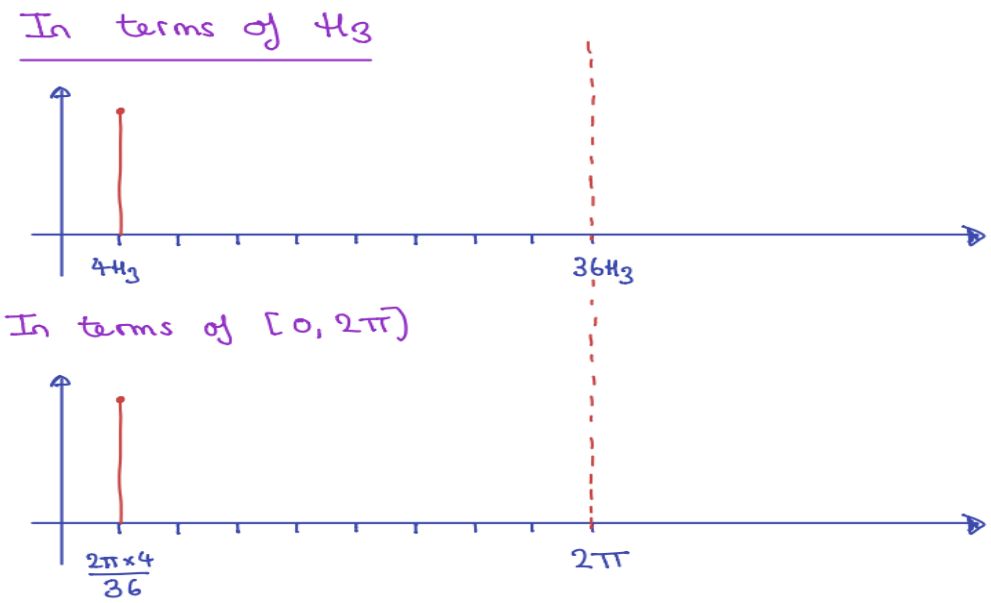
2.6.1 Connecting Hertz and Frequencies

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.



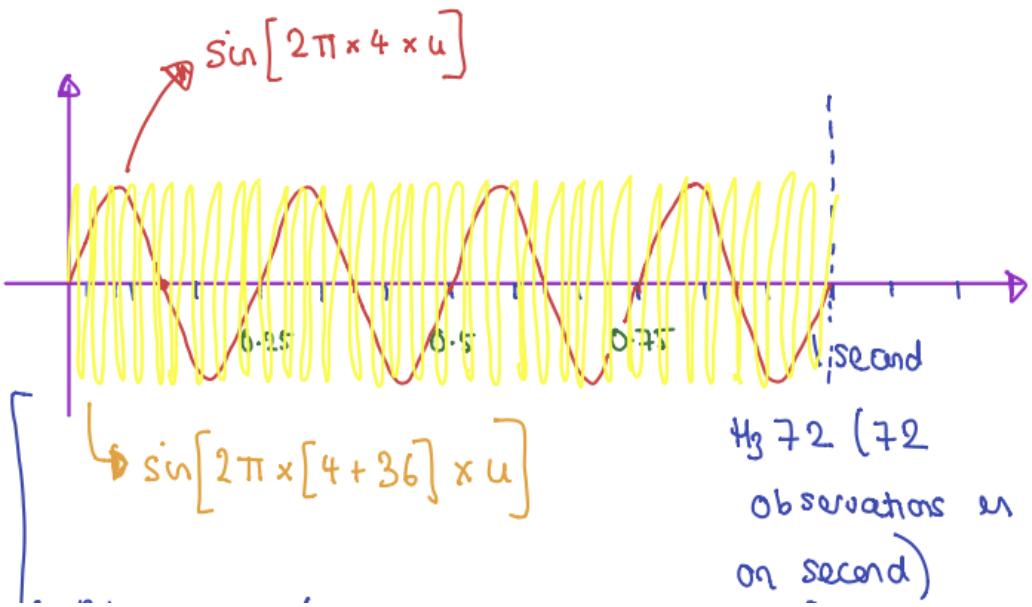
Analysis from the perspective of time series Typically, in time series, the sampling frequency is kept the same. Just the same number of second that the time series is observed grows. This allows us obtain a finer frequency grid on $[0, 2\pi]$ and obtain a better resolution in terms of peaks in frequencies. 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 $u = 1/36, 2/36, \ldots, 36/36$, then

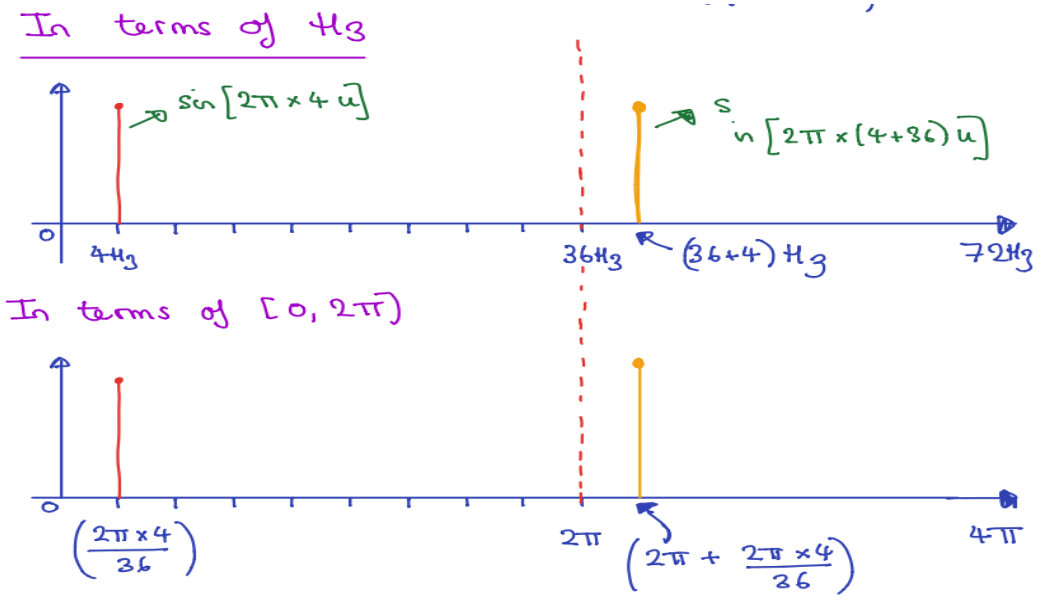
$$\sin\left(2\pi \times 4 \times \frac{t}{36}\right) = \sin\left(2\pi \times (4 + 36) \times \frac{t}{36}\right) \text{ for all } t \in \mathbb{Z}$$

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).

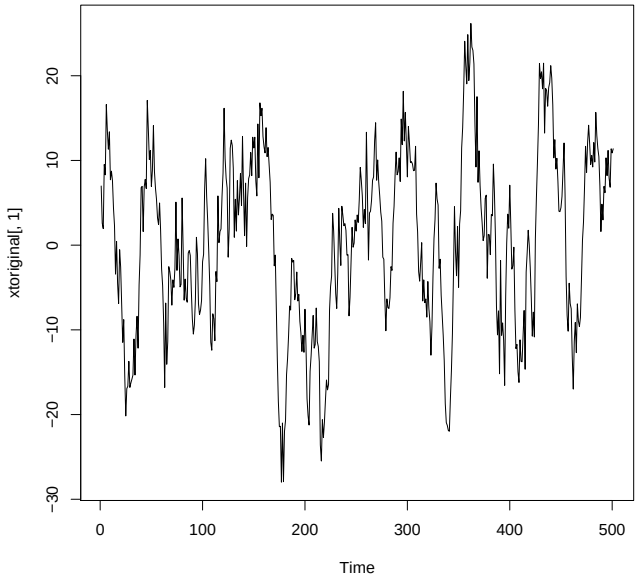


Figure 2.15: Original EEG..

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*

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

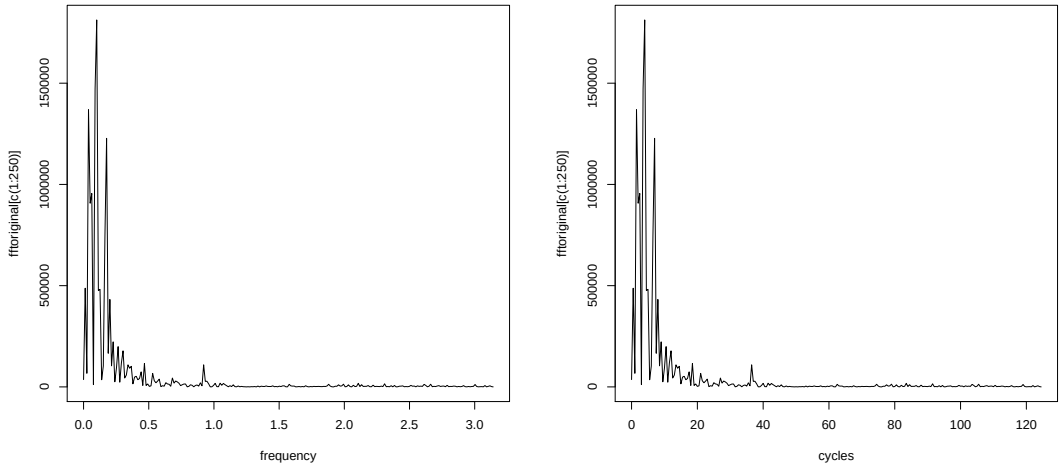


Figure 2.16: Left: Periodogram of original EEG on $[0, 2\pi]$. Right: Periodogram in terms of cycles per second.

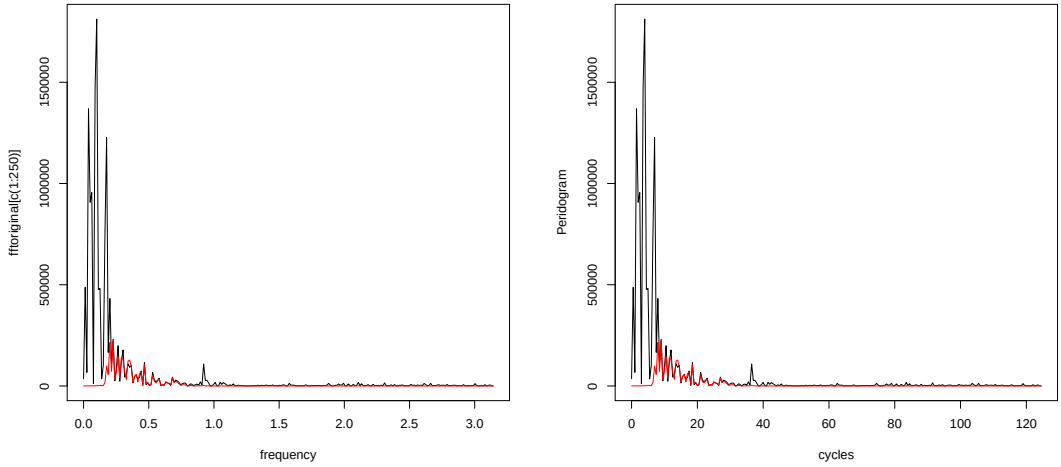


Figure 2.17: The periodogram of original EEG overlayed with processed EEG (in red). The same plot is given below, but the x-axis corresponds to cycles per second (measured in Hz)

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 $\{Y_t\}$.*

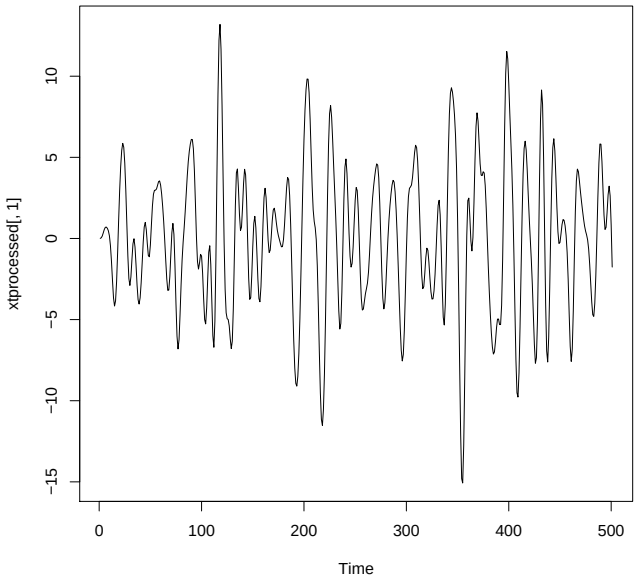


Figure 2.18: Time series after processing with a Buttersworth filter.

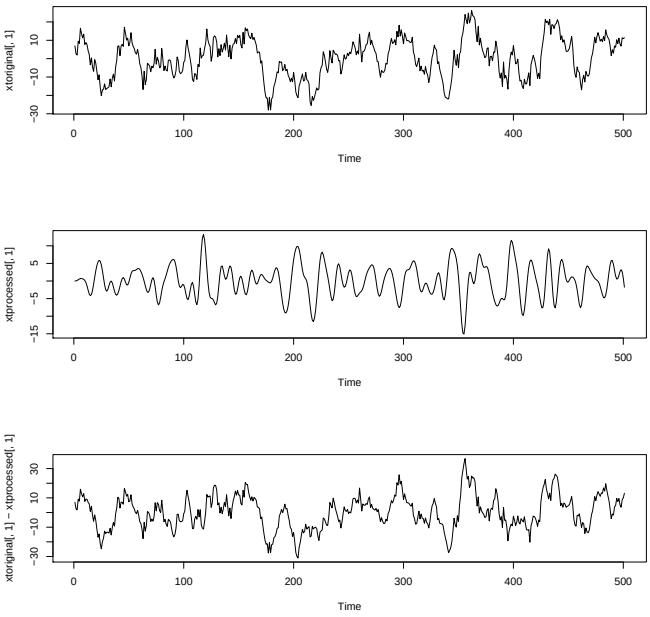


Figure 2.19: Top: Original EEG. Middle: Filtered EEG and Bottom: Difference between Original and Filtered EEG

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`*).*

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 designed only for statistics graduate students.*

(i) Let

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

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)

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

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

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

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 (2.18) using sample size $n = 60$. For each sample, estimate ω, A and B for each simulation and 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 $O(1/\sqrt{\text{number of simulations}})$.

(ii) Repeat the above experiment but this time using the sample size $n = 300$. Compare the quality 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: $X_n \stackrel{\text{a.s.}}{\to} a$ 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: $X_n \stackrel{\mathcal{P}}{\to} a$. 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 $X_n \stackrel{2}{\to} a$. 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

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

  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

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

$\beta_1 = 0$ if and only if $\text{cov}(X,Y) = 0$. 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

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

is a symmetric measure of dependence between the two variables.

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

- *Let* $X_t = \varepsilon_t + 0.5\varepsilon_{t-1}$. *Evaluate* $\text{cov}(X_t, X_{t+r})$ *for* $r = 0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5$.
- *Let* $X_t = \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j}$ *where* $|\rho| < 1$. *Evaluate* $\text{cov}(X_t, X_{t+r})$ *for* $r \in \mathbb{Z}$ *(*$0, \pm 1, \pm 2, \pm 3, \pm 4, \ldots$*).*

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

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

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

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

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

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

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

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.

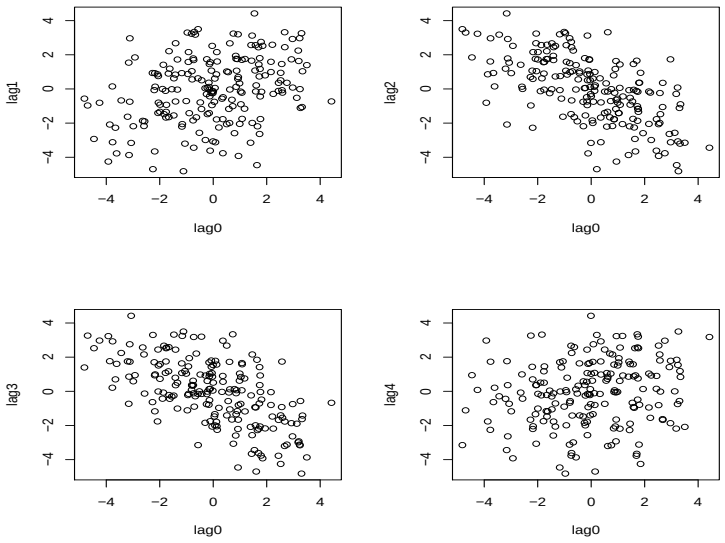


Figure 3.1: From model (3.3). Plot of X_t against X_{t+r} for $r = 1, \ldots, 4$. Top left: $r = 1$. Top right: $r = 2$, Bottom left: $r = 3$ and Bottom right: $r = 4$.

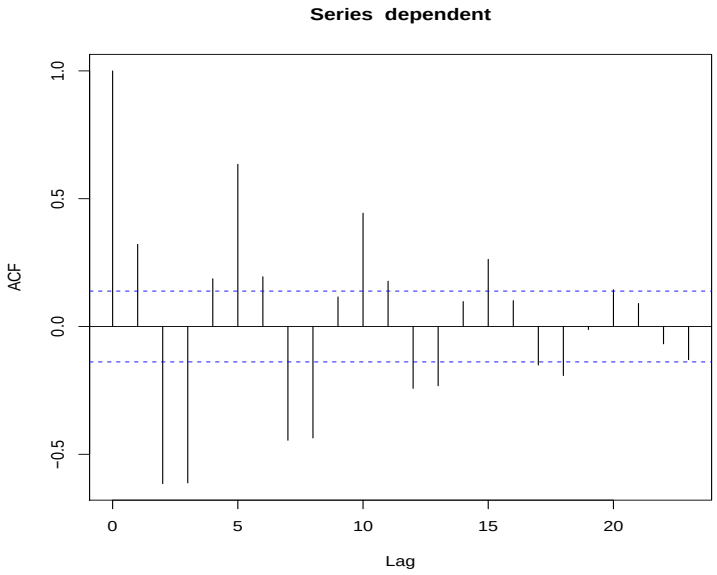


Figure 3.2: ACF plot of realisation from model (3.3).

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

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

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 $\text{var}(\bar{X})$ 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* $\text{cov}(X_t, X_{t+k})$ *we cannot say anything about* $\text{cov}(g(X_t), g(X_{t+k}))$*, 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 $\text{E}[\varepsilon_t] = 0$ 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

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

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

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

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

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

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 $\sup_t \sum_{\tau=-\infty}^{\infty} |\operatorname{cov}(\varepsilon_t, \varepsilon_\tau)| < \infty$, we can see that $\operatorname{var}\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right] = O(n^{-1})$, but just as in the case of the sample mean we need to impose some additional conditions on $\{\varepsilon_t\}$ if we want to construct confidence intervals/test $\boldsymbol{\beta}$.

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*

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

Remark 3.3.1 (Strict and second order stationarity) *(i) If a process is strictly stationarity and $\mathrm{E}|X_t^2| < \infty$, 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 $\bar{X} \stackrel{\mathcal{P}}{\to} \mu$.

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* $\mathrm{E}[V_t^2] < \infty$*.*

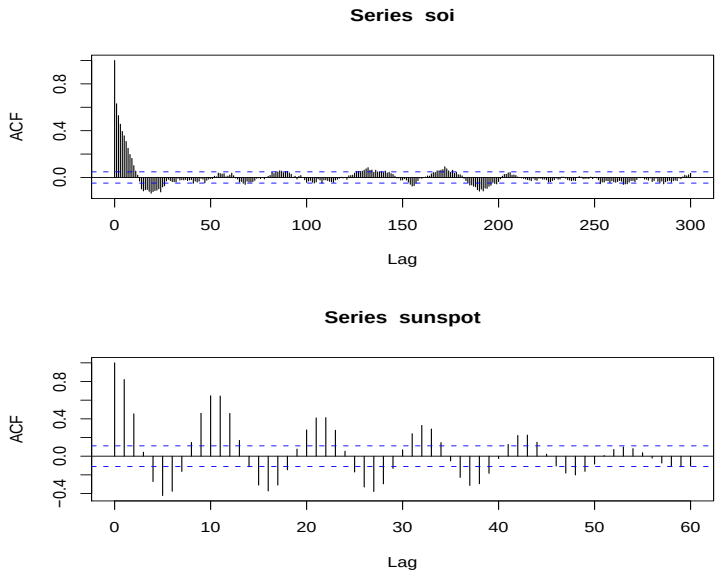


Figure 3.3: Top: ACF of Southern Oscillation data. Bottom ACF plot of Sunspot data.

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

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

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

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

(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 $\sum_{i,j}^n x_i c(i-j)x_j$. Define the random variable $Y = \sum_{i=1}^n x_i X_i$. 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 $\sum_{i,j=1}^n x_i c(i-j)x_j \geq 0$, 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

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

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

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

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

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

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 sub-

groups 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

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

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

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

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

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

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

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

(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

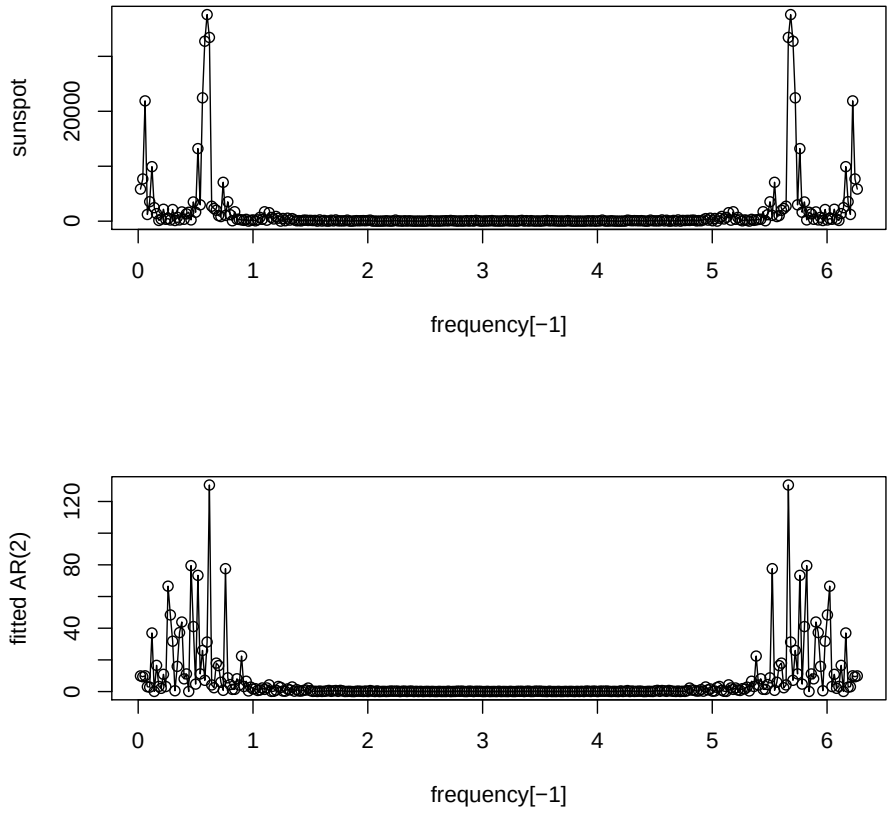


Figure 4.1: The periodogram of the Sunspot data is the top plot and the periodogram of the fitted AR(2) model is the lower plot. They do not look exactly the same, but the AR(2) model is able to capture some of the periodicities.

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.

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

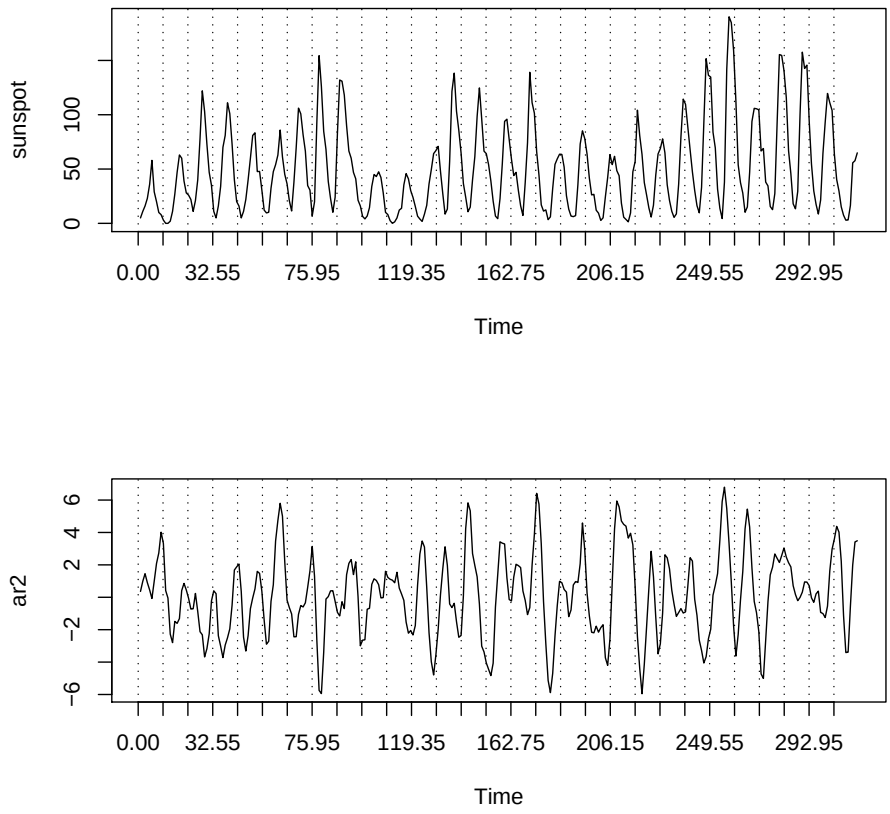


Figure 4.2: Top: Sunspot, Lower: a realisation from the AR(2) process. Lines correspond to period of $P = 2\pi/0.57 = 10.85$ years.

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

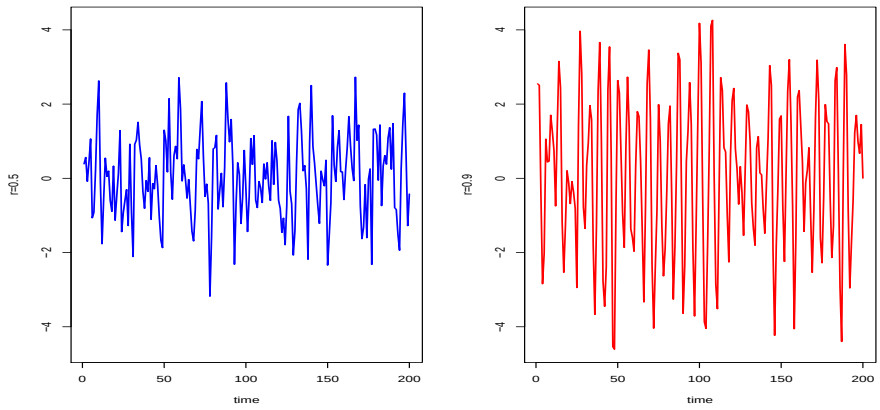


Figure 4.3: Realisation for $X_t = 2r\cos(\pi/3)X_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

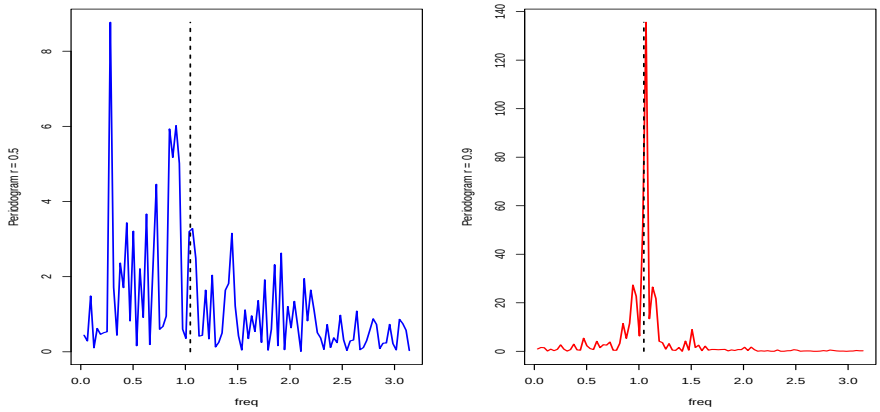


Figure 4.4: Periodogram for realisation from $X_t = 2r\cos(\pi/3)X_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

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

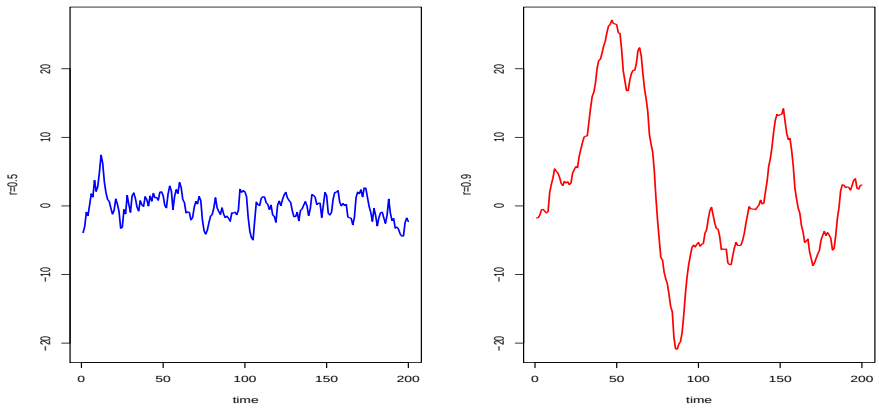


Figure 4.5: Realisation for $X_t = 2rX_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

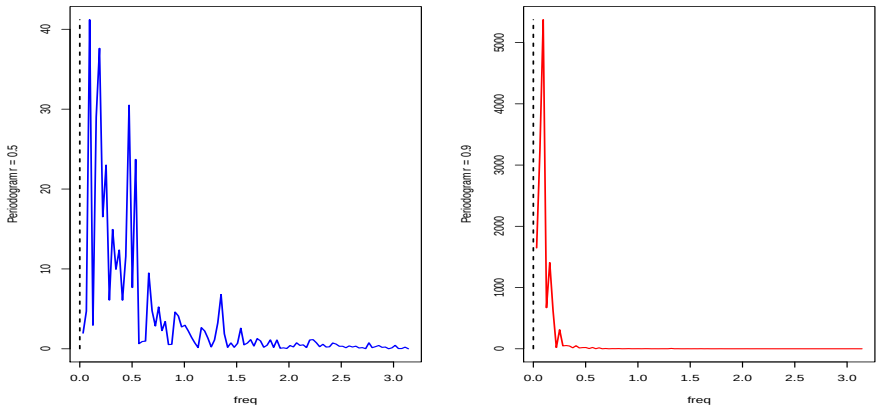


Figure 4.6: Periodogram for realisation from $X_t = 2rX_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

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.

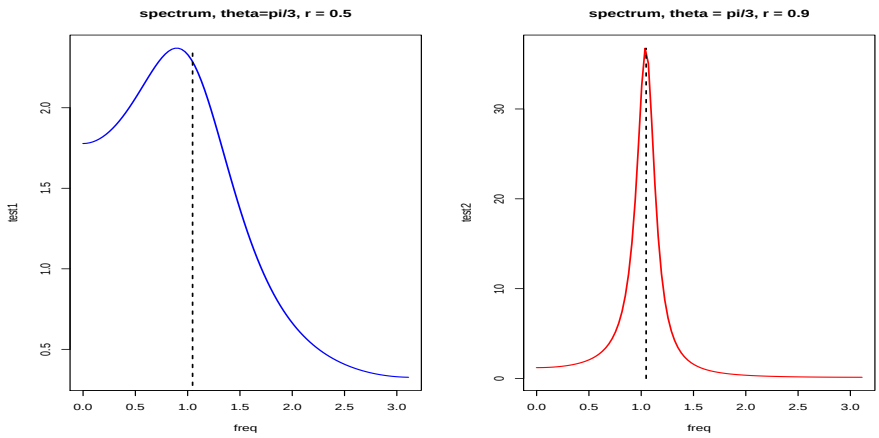


Figure 4.7: Spectral density for $X_t = 2rcos(\pi/3)X_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

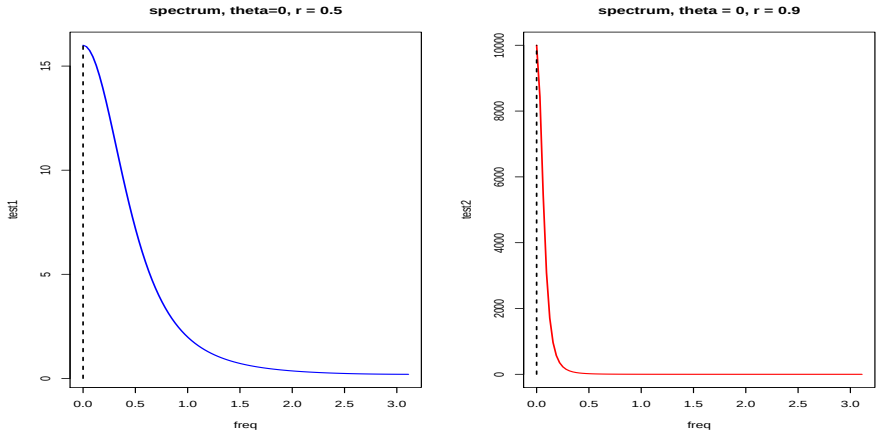


Figure 4.8: Spectral density for $X_t = 2rX_{t-1} - r^2X_{t-2} + \varepsilon_t$. Blue $= r = 0.5$ and red $= r = 0.9$.

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

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

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(∞) 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)

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

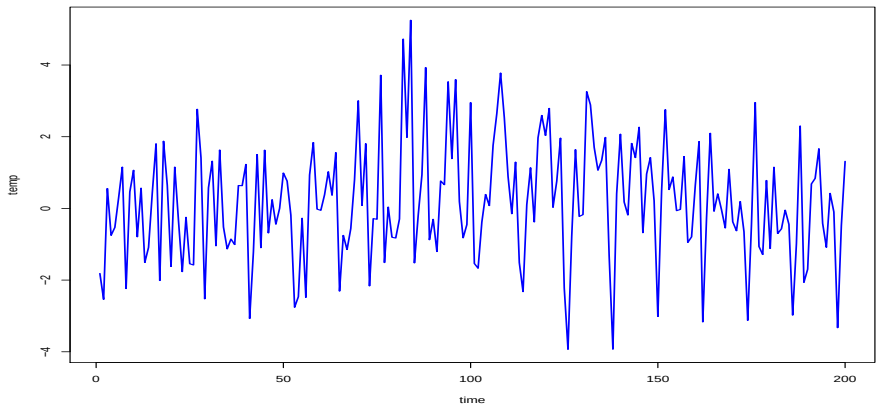


Figure 4.9: Realisation from SAR(12) with $\phi = 0.8$.

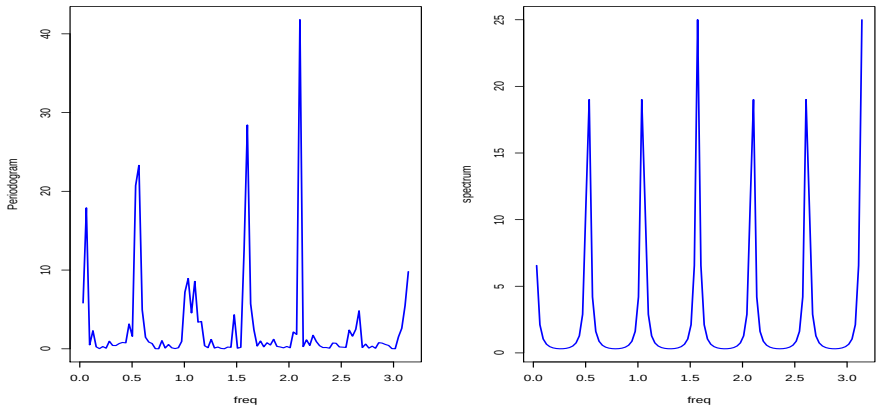


Figure 4.10: Left: Periodogram of realisation. Right Spectral density of model.

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

$$c(r) \quad = \quad \phi_1^r \sum_{j=0}^{\infty} \phi_1^{2j} = \frac{\phi_1^r}{1-\phi_1^2}.$$

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 stationary AR(4) process with characteristic function

$$\phi(z) = \left[1 - 0.8\exp(i\frac{2\pi}{13})z\right]\left[1 - 0.8\exp(-i\frac{2\pi}{13})z\right]\left[1 - \frac{2}{3}\exp(i\frac{2\pi}{5})z\right]\left[1 - \frac{2}{3}\exp(-i\frac{2\pi}{5})z\right].$$

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

$$X_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\theta_{t-q} = \varepsilon_t + \sum_{j=1}^{q}\theta_j\varepsilon_{t-j}. \tag{4.14}$$

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

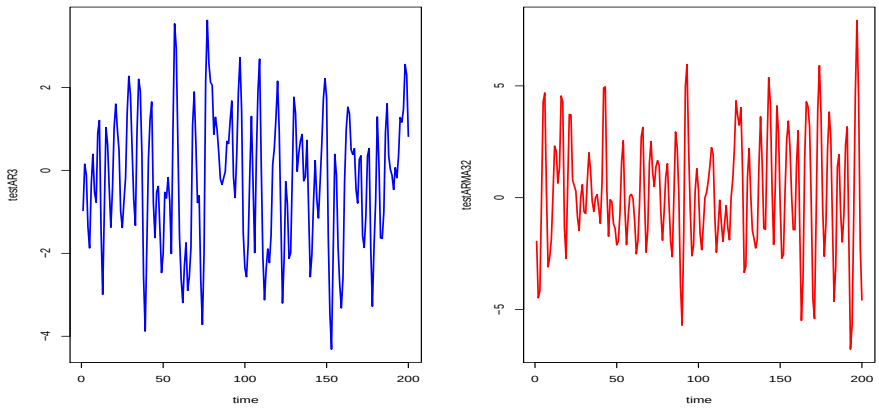


Figure 4.11: Realisation from Left: AR(3) and Right: ARMA(3, 2)

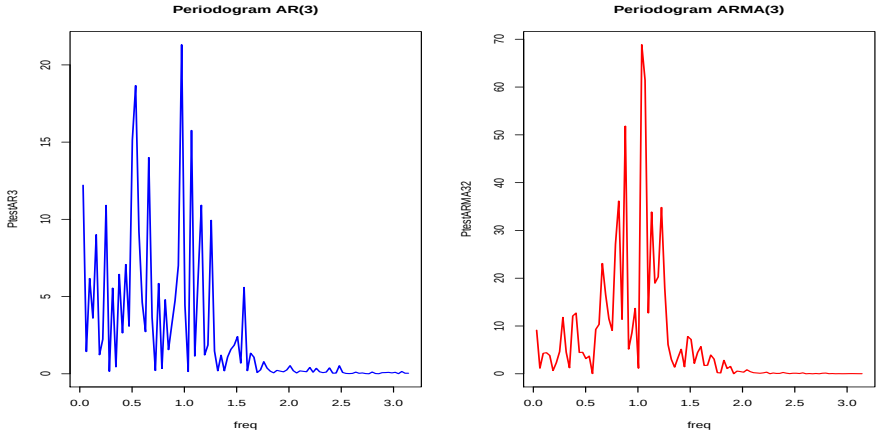


Figure 4.12: Periodogram from realisation from Left: AR(3) and Right: ARMA(3, 2)

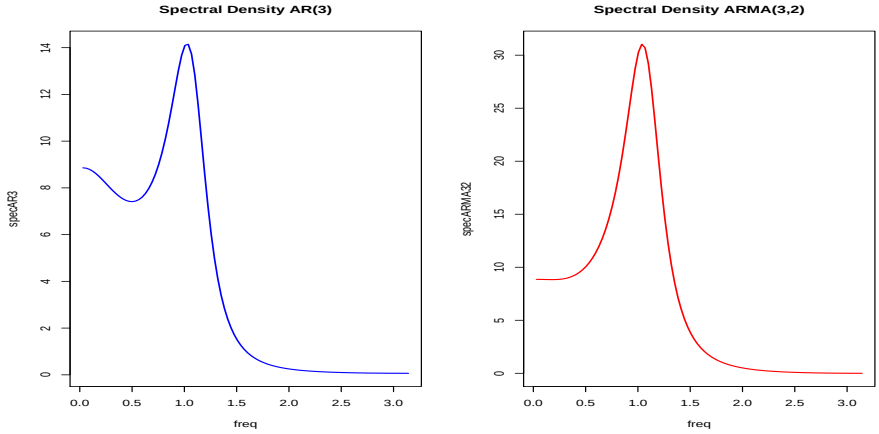


Figure 4.13: Spectral density from Left: AR(3) and Right: ARMA(3, 2)

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 $\mathrm{var}(X_t) = \mathrm{var}(X_0) + t$ (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 $(1-B)^d X_t = \varepsilon_t$ (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

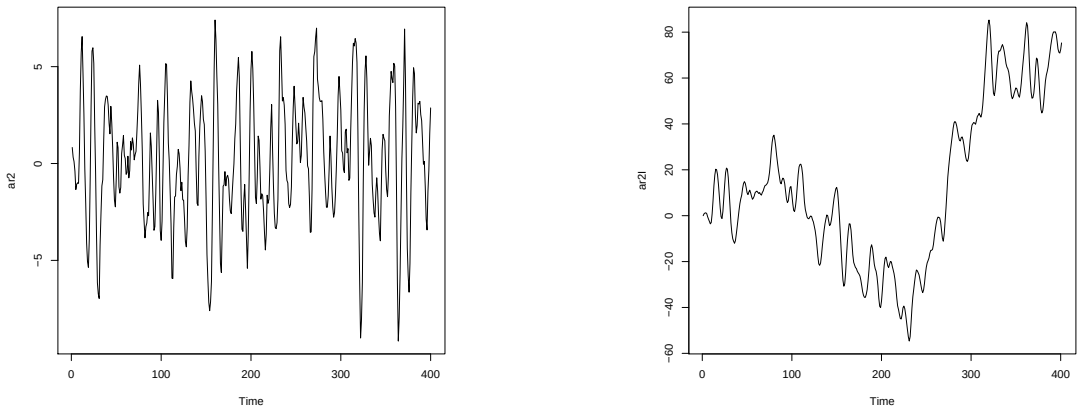
$$X_t = 2\cos\theta X_{t-1} - X_{t-2} + \varepsilon_t.$$

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).

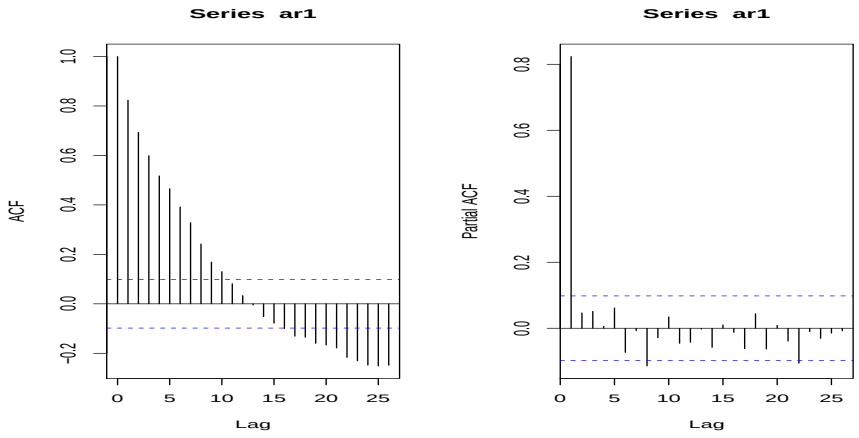


Figure 4.15: ACF and PACF plot of an AR(1), $X_t = 0.5X_{t-1} + \varepsilon_t$, $n = 400$

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

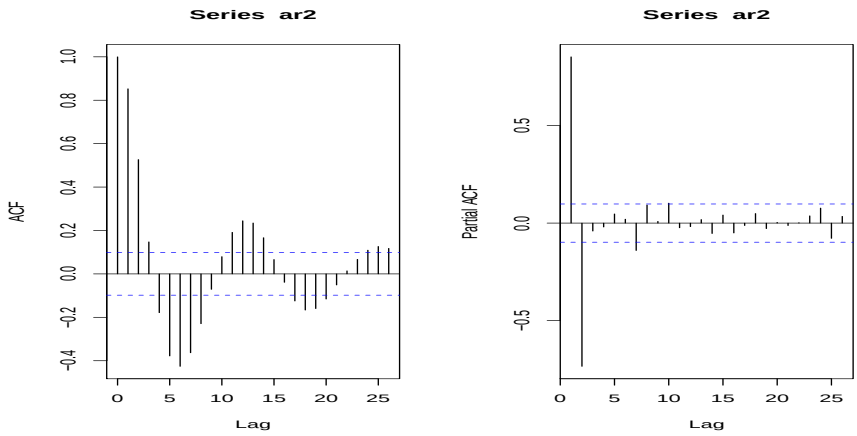


Figure 4.16: ACF and PACF plot of an AR(2), $n = 400$

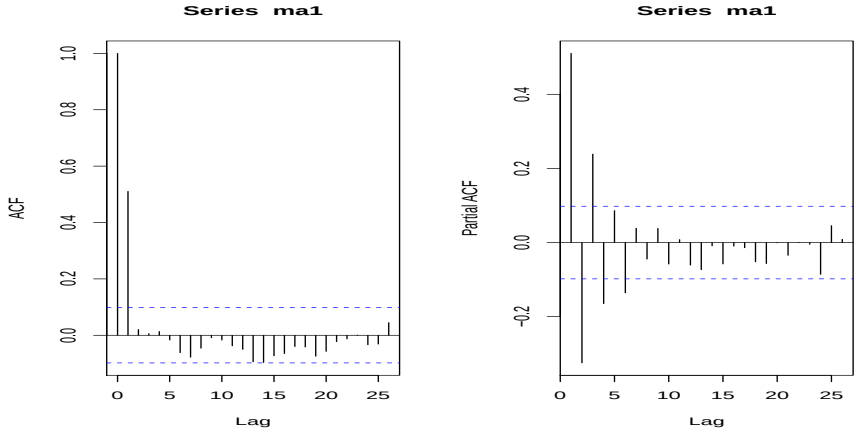


Figure 4.17: ACF and PACF plot of an MA(1), $X_t = \varepsilon_t + 0.8\varepsilon_{t-1}$, $n = 400$

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

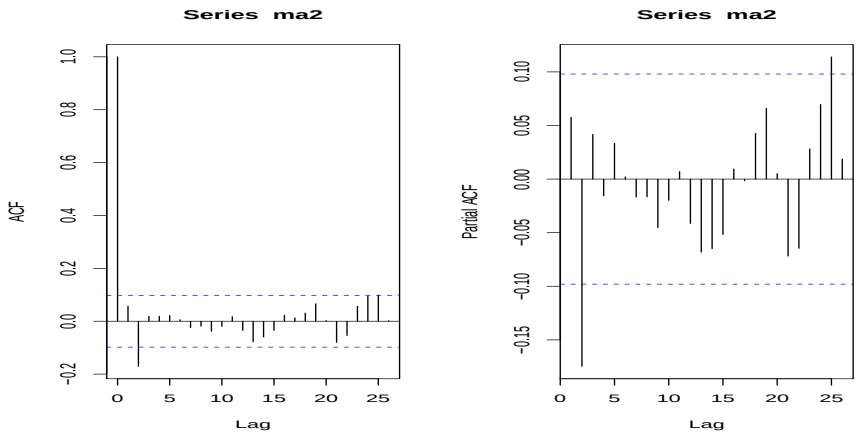


Figure 4.18: ACF and PACF plot of an MA(2), $n = 400$

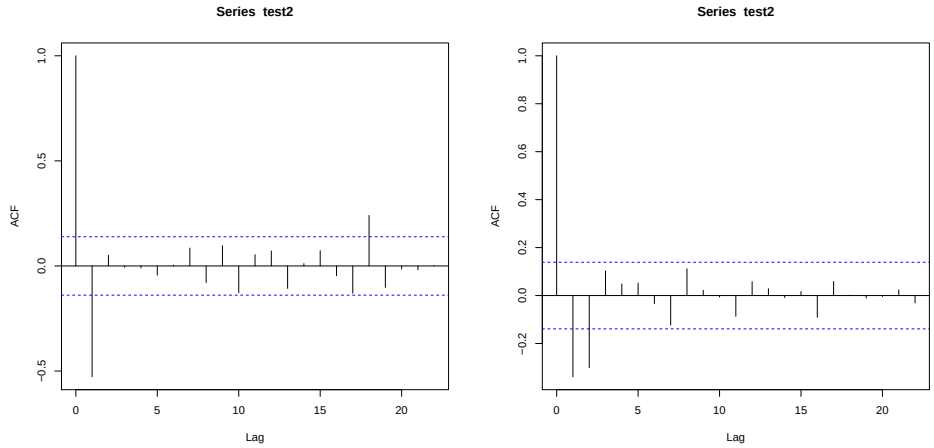


Figure 4.19: ACF of differenced data $Y_t = X_t - X_{t-1}$. Left $X_t = \varepsilon_t$, Right $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$.

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

$$\begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}, \tag{4.25}$$

$\underline{X}_t' = (X_t, \dots, X_{t-p+1})$ 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

$$\det(A - zI) = z^p - \sum_{i=1}^{p} \phi_i z^{p-i} = z^p \underbrace{(1 - \sum_{i=1}^{p} \phi_i z^{-i}))}_{=z^p\phi(z^{-1})},$$

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 $\|A^k \underline{X}_{t-k}\|_2 \leq \|A^k\|_{spec} \|\underline{X}_{t-k}\| \xrightarrow{\mathcal{P}} 0$ 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

$$(X'X) = \operatorname{diag}\left(\underline{x}_1' \underline{x}_1, \ldots, \underline{x}_p' \underline{x}_p \right).$$

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 $X_p = \operatorname{sp}(\underline{x}_1, \ldots, \underline{x}_p)$ and $X_{p+1} = \operatorname{sp}(\underline{x}_1, \ldots, \underline{x}_{p+1})$. Observe that X_p is a subset of X_{p+1}. With a little thought it is clear that $X_{p+1} = \operatorname{sp}(X_p, \underline{x}_{p+1} - P_{X_p}(\underline{x}_{p+1}))$. 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

$$H = \{X; \text{X is a (real) random variables where } \text{var}(X) < \infty\}.$$

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

$$\langle X, Y \rangle = \text{cov}(X, Y).$$

The norm is clearly the variance

$$\|X\|_2^2 = \langle X, X \rangle = \text{cov}(X, X).$$

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 $\text{sp}(X_1, \ldots, X_n)$

$$\text{sp}(X_1, \ldots, X_n) = \left\{ Y; \text{where } Y = \sum_{j=1}^{p} a_j X_j \right\},$$

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 $X = \text{sp}(X_1, \ldots, X_n)$. 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

$$\langle X_\ell, Y - \sum_{j=1}^{p} \alpha_j X_j \rangle = \text{cov}(X_\ell, Y - \sum_{j=1}^{p} \alpha_j X_j) = 0 \qquad 1 \leq j \leq p.$$

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

$$\mathrm{E}\left(Y - \sum_{j=1}^{p} \beta_j X_j\right)^2 \quad = \quad \mathrm{E}(Y^2) - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.$$

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

$$\mathrm{E}\left(Y - \sum_{j=1}^{p}\beta_j X_j\right)^2 \quad = \quad \mathrm{E}(Y^2) - 2\mathrm{E}(Y\widehat{Y}) + \mathrm{E}(\widehat{Y}^2).$$

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

$$\begin{aligned}\mathrm{E}\left(Y - \sum_{j=1}^{p}\beta_j X_j\right)^2 &= \mathrm{E}(Y^2) - \mathrm{E}(\widehat{Y}^2) = \mathrm{E}(Y^2) - \underline{\beta}'\mathrm{var}(\mathbf{X})\underline{\beta} \\ &= \mathrm{E}(Y^2) - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.\end{aligned}$$

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

$$P_Y(X) = \underline{\alpha}'_{X,Y}\underline{Y} \text{ and } P_Y(Z) = \underline{\alpha}'_{Z,Y}\underline{Y}$$

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.

$$\begin{aligned}
\mathrm{cov}(X - P_Y(X), Z - P_Y(Z)) &= \mathrm{cov}(X,Z) - \mathrm{cov}(X,Y)'[\mathrm{var}(Y)]^{-1}\mathrm{cov}(Z,Y) \\
\mathrm{var}(X - P_Y(X)) &= \mathrm{var}(X) - \mathrm{cov}(X,Y)'[\mathrm{var}(Y)]^{-1}\mathrm{cov}(X,Y) \\
\mathrm{var}(Z - P_Y(Z)) &= \mathrm{var}(Z) - \mathrm{cov}(Z,Y)'[\mathrm{var}(Y)]^{-1}\mathrm{cov}(Z,Y).
\end{aligned}$$

Relating partial correlation and the regression cofficients 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

$$\beta_{Z\to X} = \frac{\text{cov}(X, Z - P_Y(Z))}{\text{var}(Z - P_Y(Z))} \text{ and } \beta_{X\to Z} = \frac{\text{cov}(Z, X - P_Y(X))}{\text{var}(X - P_Y(X))}.$$

Since $Z - P_Y(Z)$ is orthogonal to Y (and thus $\text{cov}(Z - P_Y(Z), P_Y(X)) = 0$) we have

$$\text{cov}(X, Z - P_Y(Z)) = \text{cov}(X - P_Y(X), Z - P_Y(Z)).$$

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

$$\beta_{Z\to X} = \frac{\text{cov}(X - P_Y(X), Z - P_Y(Z))}{\text{var}(Z - P_Y(Z))} \text{ and } \beta_{X\to Z} = \frac{\text{cov}(Z - P_Y(Z), X - P_Y(X))}{\text{var}(X - P_Y(X))}.$$

To connect the regression coefficients to the partial correlations we rewrite we rewrite the partial covariance in terms of the partial correlation:

$$\text{cov}(X - P_Y(X), Z - P_Y(Z)) = \rho_{X,Z|Y}\sqrt{\text{var}(X - P_Y(X))\text{var}(Z - P_Y(Z))}.$$

Substituting the expression for $\text{cov}(X - P_Y(X), Z - P_Y(Z))$ into the expression for $\beta_{Z\to X}$ and $\beta_{X\to Z}$ gives

$$\beta_{Z\to X} = \rho_{X,Z|Y}\sqrt{\frac{\text{var}(X - P_Y(X))}{\text{var}(Z - P_Y(Z))}} \text{ and } \beta_{X\to Z} = \rho_{X,Z|Y}\sqrt{\frac{\text{var}(Z - P_Y(Z))}{\text{var}(X - P_Y(X))}}. \tag{5.9}$$

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 $\text{var}[\varepsilon_X] = \text{var}[X - P_{Y,Z}(X)]$ and $\text{var}[\varepsilon_Z] = \text{var}[Z - P_{Y,X}(Z)]$.

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 $\text{var}[\varepsilon_X] = \text{var}[X - P_{Y,Z}(X)]$ and $\text{var}[\varepsilon_Z] = \text{var}[Z - P_{Y,X}(Z)]$.

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

$$\text{var}(\varepsilon_{j,-i})\text{var}(\varepsilon_{i,-j}) - \text{var}(\varepsilon_j)\text{var}(\varepsilon_{i,-j}) = \text{var}(\varepsilon_{i,-j})\text{var}(\varepsilon_{j,-i}) - \text{var}(\varepsilon_i)\text{var}(\varepsilon_{j,-i}).$$

This leads to the required identity

$$\frac{\text{var}(\varepsilon_i)}{\text{var}(\varepsilon_j)} = \frac{\text{var}(\varepsilon_{i,-j})}{\text{var}(\varepsilon_{j,-i})},$$

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 $\text{cov}(X_1, X_2) = \text{var}(X_3)$ 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

$$\Sigma = \text{var}[\boldsymbol{X}] \text{ and } \Gamma = \Sigma^{-1}.$$

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 $\text{E}[X_i - \boldsymbol{\beta}_i'\mathbf{X}_{-i}]^2$, 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

$$\begin{aligned} \mathrm{cov}\left(\varepsilon_{i,-j}, \varepsilon_{j,-i}\right) &= \mathrm{cov}(X_i, X_j) - \mathrm{cov}(X_i, \mathbf{X}'_{-(i,j)})\mathrm{var}[\mathbf{X}_{-(i,j)}]^{-1}\mathrm{cov}(X_j, \mathbf{X}_{-(i,j)}) \\ \mathrm{var}\left(\varepsilon_{i,-j}\right) &= \mathrm{var}(X_i) - \mathrm{cov}(X_i, \mathbf{X}'_{-(i,j)})\mathrm{var}[\mathbf{X}_{-(i,j)}]^{-1}\mathrm{cov}(X_i, \mathbf{X}_{-(i,j)}) \\ \mathrm{var}\left(\varepsilon_{j,-i}\right) &= \mathrm{var}(X_j) - \mathrm{cov}(X_j, \mathbf{X}'_{-(i,j)})\mathrm{var}[\mathbf{X}_{-(i,j)}]^{-1}\mathrm{cov}(X_j, \mathbf{X}_{-(i,j)}). \end{aligned}$$

- 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

$$\Big(\; -\beta_{i1}/\sigma_i^2, \quad -\beta_{i2}/\sigma_i^2, \quad \ldots \quad 1/\sigma_i^2 \quad \ldots \quad -\beta_{id}/\sigma_i^2. \; \Big)$$

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

$$\Big(\; -\beta_{i1}/\sigma_i^2 \quad -\beta_{i2}/\sigma_i^2 \quad \ldots \quad -\beta_{id}/\sigma_i^2. \; \Big).$$

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

$$\rho_{ij} = \beta_{ij} \sqrt{\frac{\text{var}[\varepsilon_j]}{\text{var}[\varepsilon_i]}}. \tag{5.14}$$

We recall that $\Gamma_{ii} = \text{var}(X_i - P_{X_{-i}}(X_i))^{-1}$, $\Gamma_{jj} = \text{var}(X_j - P_{X_{-j}}(X_j))^{-1}$ and $\Gamma_{ij} = -\beta_{ij}\Gamma_{ii}$ gives

$$\rho_{ij} = -\frac{\Gamma_{ij}}{\Gamma_{ii}} \sqrt{\frac{\Gamma_{ii}}{\Gamma_{jj}}} = -\frac{\Gamma_{ij}}{\sqrt{\Gamma_{ii}\Gamma_{jj}}}.$$

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 $\text{cov}(X_i, X_j) = \Sigma_{ij}$. 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

$$\Sigma^{-1} = \begin{pmatrix} P^{-1} & -P^{-1}\underline{c}_{1,2}'\Sigma_{-(1,2)}^{-1} \\ -\Sigma_{-(1,2)}^{-1}\underline{c}_{1,2}P^{-1} & P^{-1} + \Sigma_{-(1,2)}^{-1}\underline{c}_{1,2}P^{-1}\underline{c}_{1,2}'\Sigma_{-(1,2)}^{-1} \end{pmatrix}, \tag{5.18}$$

where $P = (\Sigma_{1,2} - \underline{c}_{1,2}'\Sigma_{-(1,2)}^{-1}\underline{c}_{1,2})$. 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

$$L = \begin{pmatrix} \gamma_{1,1} & 0 & 0 & \ldots & 0 & 0 \\ \gamma_{2,1} & \gamma_{2,2} & 0 & \ldots & 0 & 0 \\ \gamma_{3,1} & \gamma_{3,2} & \gamma_{3,3} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{k,1} & \gamma_{k,2} & \gamma_{k,3} & \cdots & \gamma_{k,k-1} & \gamma_{k,k} \end{pmatrix} \tag{5.32}$$

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 $\Sigma = \text{var}(\underline{X})$ (where $\underline{X}$ is a p-dimension random vector), then*

$$\text{var}\,(L\underline{X}) = L'\Sigma L = L'H'HL = I_p.$$

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. $\text{E}(X^2) = \int_\Omega X(\omega)^2 dP(\omega) < \infty$. This space is denoted as $L^2(\Omega, \mathcal{F}, P)$. In this case, the inner product is $\langle X, Y\rangle = \text{E}(XY)$.

(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

$$Y = \sum_{j=1}^{t} a_j X_{t+1-j}. \tag{5.33}$$

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

$$X_t = 0.5 X_{t-1} + \varepsilon_t$$

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 same autocorrelation function. They are duals. In summary, autocorrelation is 'blind' to non-causality.

Another worked example Consider the AR(2) model

$$X_t = 2r\cos(\theta)X_{t-1} - r^2X_{t-2} + \varepsilon_t, \tag{6.3}$$

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

$$c(k) = ar^{|k|}\left(\exp(i(b+k\theta))+\exp(-i(b+k\theta))\right) = 2ar^{|k|}\cos\left(k\theta+b\right), \tag{6.4}$$

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

$$X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t, \tag{6.5}$$

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

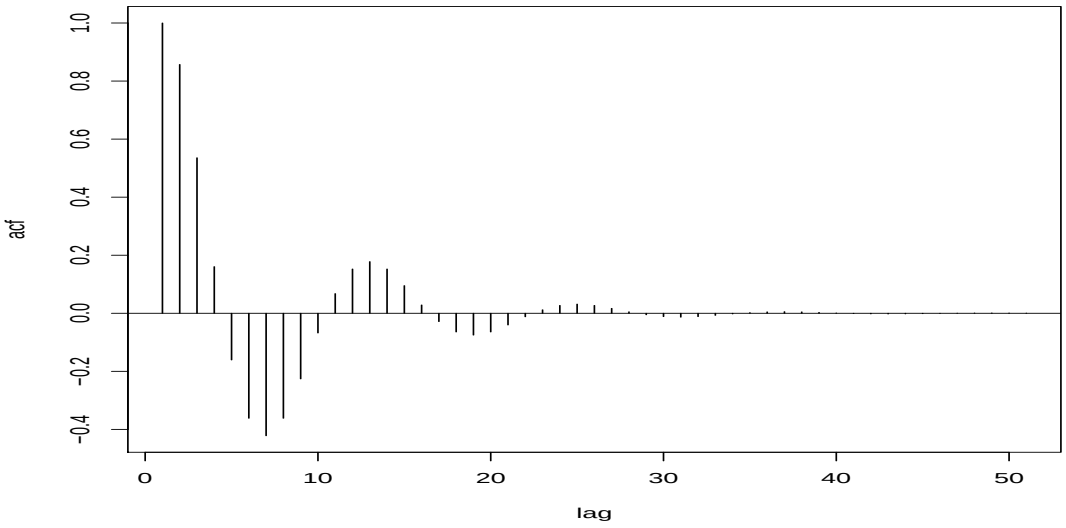


Figure 6.1: The ACF of the time series $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$

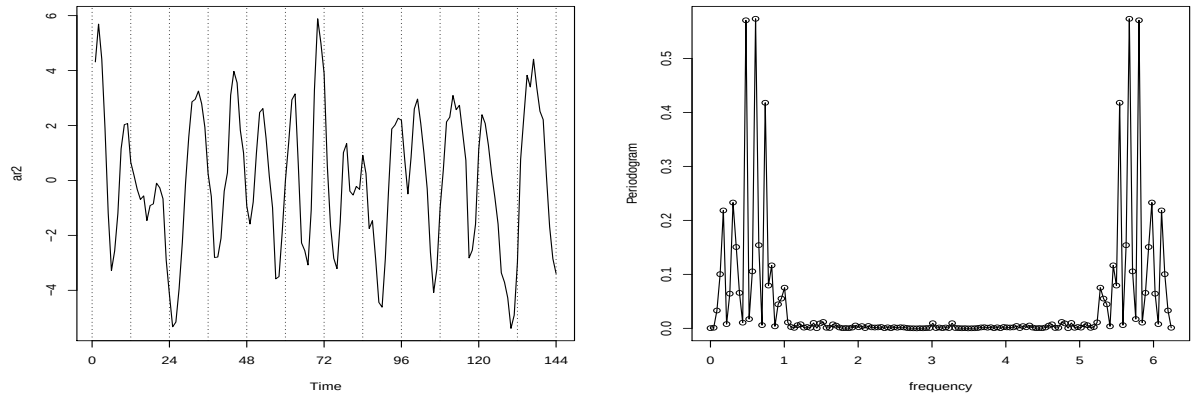


Figure 6.2: Left: A realisation from the time series $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$. Right: The corresponding periodogram.

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

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

By matching the innovations it can be shown that for $r > 0$

$$\mathrm{cov}(X_t, X_{t+r}) \;=\; \sum_{j=0}^{\infty}(b^{j+1} - a^{j+1})(b^{j+1+r} - a^{j+1+r}). \tag{6.6}$$

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

$$\mathrm{cov}(X_t X_{t-k}) \;=\; \sum_{j=1}^{p}\phi_j \mathrm{cov}(X_{t-j}, X_{t-k}). \tag{6.8}$$

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 $c(r) = \mathrm{cov}(X_0, X_r)$ 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

$$c(k) \;\; = \;\; C_1 \lambda_1^{-|k|} + C_2 \lambda_2^{-|k|} = C\lambda^{-|k|} + \overline{C\lambda}^{-|k|}. \tag{6.11}$$

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

$$\mathrm{cov}(X_t, X_{t-k}) - \sum_{i=1}^{p} \phi_i \mathrm{cov}(X_{t-i}, X_{t-k}) \;=\; \sum_{j=1}^{q} \theta_j \mathrm{cov}(\varepsilon_{t-j}, X_{t-k}) = \sum_{j=k}^{q} \theta_j \mathrm{cov}(\varepsilon_{t-j}, X_{t-k}).$$

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 $\mathrm{cov}(\varepsilon_{t-j}, X_{t-k}) = a_{j-k}\mathrm{var}(\varepsilon_t)$ (where $a(z) = \theta(z)\phi(z)^{-1}$). Altogether the above gives the difference equations

$$\begin{aligned} c(k) - \sum_{i=1}^{p} \phi_i c(k-i) &= \mathrm{var}(\varepsilon_t) \sum_{j=k}^{q} \theta_j a_{j-k} \quad \text{for } 1 \leq k \leq q \\ c(k) - \sum_{i=1}^{p} \phi_i c(k-i) &= 0, \text{ for } k > q, \end{aligned}$$

where $c(k) = \mathrm{cov}(X_0, X_k)$. 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 $c(k) = \mathrm{cov}(Y_0, Y_k)$ 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 $h \geq 1$

$$\sqrt{n}\widehat{c}_n(h) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1). \tag{6.14}$$

This gives rise to the critical values $\pm 1.96/\sqrt{n}$.

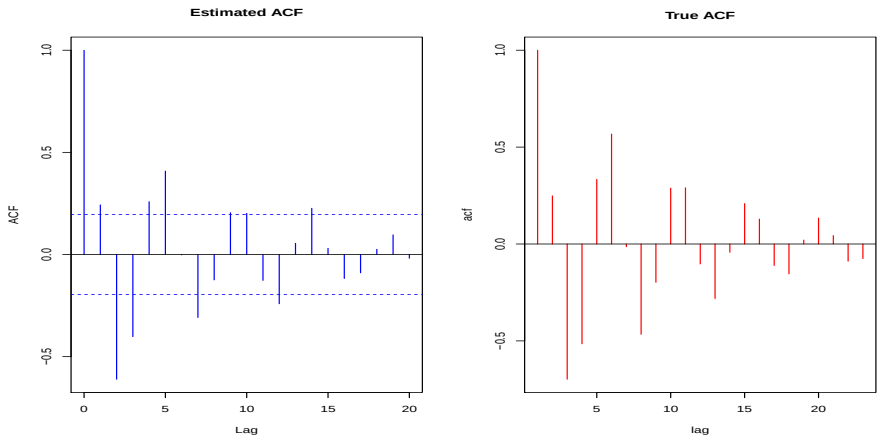


Figure 6.3: The AR(2) model. Left: Estimated ACF based on $n = 100$. Right: True ACF

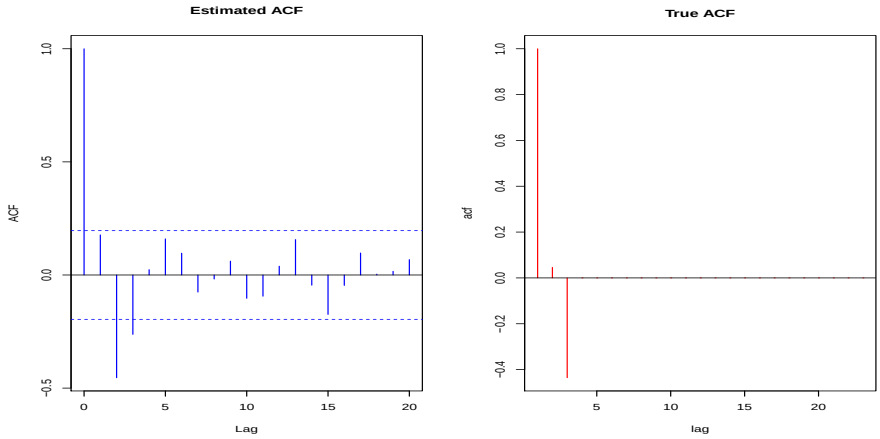


Figure 6.4: The MA(2) model. Left: Estimated ACF based on $n = 100$. Right: True ACF

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

$$\rho_k(t) = \frac{\mathrm{cov}(X_t - P_{\underline{Y}}(X_t), X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1}))}{\sqrt{\mathrm{var}(X_t - P_{\underline{Y}}(X_t))\mathrm{var}(X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1}))}},$$

with

$$\begin{aligned}
& \mathrm{cov}(X_t - P_{\underline{Y}}(X_t), X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1})) \\
= \; & \mathrm{cov}(X_t, X_{t+k+1}) - \mathrm{cov}(X_t, \underline{Y})'[\mathrm{var}(\underline{Y})]^{-1}\mathrm{cov}(X_{t+k+1}, \underline{Y}) \\
& \mathrm{var}(X_t - P_{\underline{Y}}(X_t)) \\
= \; & \mathrm{var}(X_t) - \mathrm{cov}(X_t, \underline{Y})'[\mathrm{var}(\underline{Y})]^{-1}\mathrm{cov}(X_t, \underline{Y}) \\
& \mathrm{var}(X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1})) \\
= \; & \mathrm{var}(X_{t+k+1}) - \mathrm{cov}(X_{t+k+1}, \underline{Y})'[\mathrm{var}(\underline{Y})]^{-1}\mathrm{cov}(X_{t+k+1}, \underline{Y}).
\end{aligned}$$

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 $\mathrm{cov}(X_t, X_{t+k+1}) = c(k+1)$, $\mathrm{cov}(X_t, X_t) = c(0)$ etc). We can center everything about $t = 0$, the only term that is relevant is the spacing k and define

$$\rho_{k+1|k+1} = \frac{\mathrm{cov}(X_0 - P_{\underline{Y}}(X_0), X_{k+1} - P_{\underline{Y}}(X_{k+1}))}{\sqrt{\mathrm{var}(X_0 - P_{\underline{Y}}(X_0))\mathrm{var}(X_{k+1} - P_{\underline{Y}}(X_{k+1}))}},$$

where $\underline{Y}' = (X_1, X_2, \ldots, X_k)$,

$$\begin{aligned}
\mathrm{cov}(X_t - P_{\underline{Y}}(X_t), X_{t+k+1} - P_{\underline{Y}}(X_{t+k+1})) &= c(k+1) - \mathrm{cov}(X_0, \underline{Y})'[\mathrm{var}(\underline{Y})]^{-1}\mathrm{cov}(X_{k+1}, \underline{Y}) \\
\mathrm{var}(X_0 - P_{\underline{Y}}(X_0)) &= c(0) - \mathrm{cov}(X_0, \underline{Y})'[\mathrm{var}(\underline{Y})]^{-1}\mathrm{cov}(X_0, \underline{Y}) \\
\mathrm{var}(X_{k+1} - P_{\underline{Y}}(X_{k+1})) &= c(0) - \mathrm{cov}(X_{k+1}, \underline{Y})'[\mathrm{var}(\underline{Y})]^{-1}\mathrm{cov}(X_{k+1}, \underline{Y}).
\end{aligned}$$

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

$$\begin{aligned} \text{cov}(X_0, \underline{Y}') &= (c(1), c(2) \ldots, c(k+1)) \\ \text{and } \text{cov}(X_{n+1}, \underline{Y}') &= (c(k+1), c(2) \ldots, c(1)). \end{aligned}$$

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.

Time reversibility property of stationary time series For stationary time series, predicting into the future and predicting into the past leads to the same set 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:

$$P_{\underline{Y}}(X_0) = \sum_{j=1}^{k} \phi_{k,j} X_j = \sum_{j=1}^{k} \phi_{k,k+1-j} X_{k+1-j}. \tag{6.15}$$

Proof of equation (6.15)

$$P_{\underline{Y}}(X_0) = \underline{Y}' (\text{var}[\underline{Y}]^{-1} \text{cov}[X_0, \underline{Y}]).$$

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

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t \tag{6.17}$$

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:

$$\rho_{p+1|p+1} = \phi_{p+1|p+1}\sqrt{\frac{\text{var}(\varepsilon_{0|X_1,\ldots,X_{p+1}})}{\text{var}(\varepsilon_{p+1|X_0,\ldots,X_p})}}$$

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 $\text{var}(\varepsilon_{0|X_1,\ldots,X_{p+1}}) = \text{var}(\varepsilon_{p+1|X_0,\ldots,X_p})$ 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.

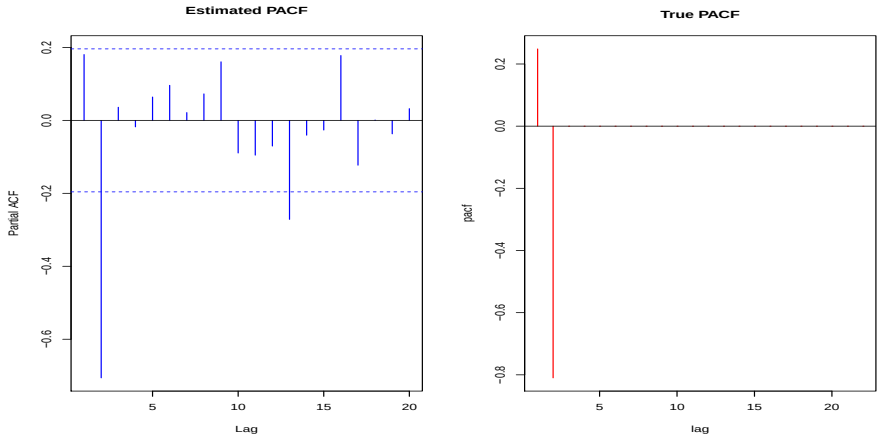


Figure 6.5: The AR(2): Left Estimated PACF ($n = 100$). Right: True PACF plot. $n = 100$

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).

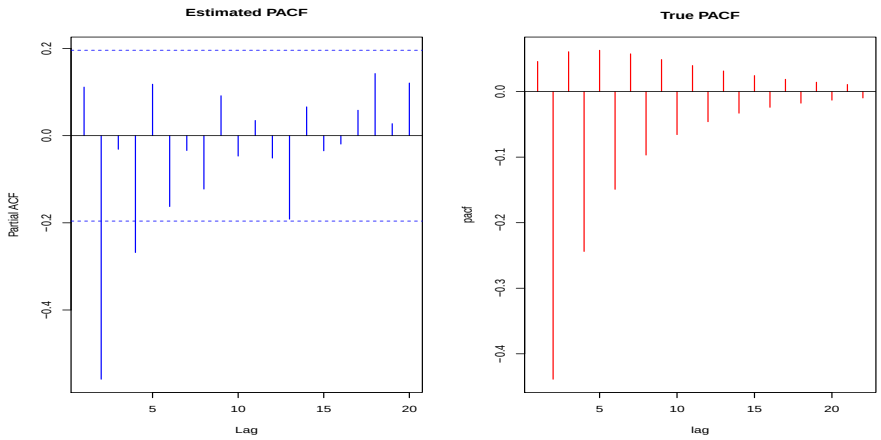


Figure 6.6: The MA(2): Left Estimated PACF ($n = 100$). Right: True PACF plot. $n = 100$

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 $h > p$

$$\sqrt{n}\widehat{\rho}_{h|h} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \tag{6.19}$$

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 $\text{var}(\underline{X}_n) = \Sigma_k$, 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

$$\Sigma_n = \text{var}(\boldsymbol{X}_n) = \begin{pmatrix} c(0) & c(1) & 0 & \ldots & c(n-2) & c(n-1) \\ c(1) & c(0) & c(1) & \ldots & c(n-3) & c(n-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ c(n-1) & c(n-2) & \vdots & \ldots & c(1) & c(0) \end{pmatrix}.$$

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

$$L_n = \begin{pmatrix} \gamma_{1,0} & 0 & \dots & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ -\gamma_{2,1} & \gamma_{2,0} & \dots & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\phi_p & -\phi_{p-1} & \dots & -\phi_1 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & -\phi_p & -\phi_{p-1} & \dots & -\phi_1 & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & -\phi_p & \dots & -\phi_2 & -\phi_1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \tag{6.21}$$

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 $\underline{X}_n' = (X_1, \dots, X_n)$ 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 $\text{var}(\varepsilon_t) < \infty$. 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

$$\text{E}(X_t X_{t-k}) = \sum_{j=1}^{p} \phi_j \text{E}(X_{t-j} X_{t-k}) \Rightarrow c(k) - \sum_{j=1}^{p} \phi_j c(k-j) = 0. \tag{6.28}$$

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 $\mathrm{var}(X_t|X_{t-1},\ldots,X_1)$ 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

$$\mathrm{E}[X_{t+r} - X_t(r)]^2 = \mathrm{E}[\sum_{i=0}^{r-1} \phi^i \varepsilon_{t+r-i}]^2 = \mathrm{var}[\varepsilon_t] \sum_{i=0}^{r-1} \phi^{2i}.$$

Worked example: AR(2) 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)

$$\begin{pmatrix}X_{t+1}\\X_t\\\vdots\\\vdots\\X_{t-p+1}\end{pmatrix}=\begin{pmatrix}\phi_1&\phi_2&\phi_3&\ldots&\phi_p\\1&0&0&\ldots&0\\0&1&0&\ldots&0\\\vdots&\ldots&\vdots&\ddots&0\\0&0&0&\ldots&0\end{pmatrix}\begin{pmatrix}X_t\\X_{t-1}\\\vdots\\\vdots\\X_{t-p}\end{pmatrix}+\begin{pmatrix}\varepsilon_{t+1}\\0\\0\\\vdots\\0\end{pmatrix}$$

$$\Rightarrow \underline{X}_{t+1} = \Phi\underline{X}_t+\underline{\varepsilon}_{t+1}.$$

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

$$\begin{aligned} X_t(r) &= \sum_{j=0}^{\infty} \psi_{j+r}\varepsilon_{t-j} \\ &= \sum_{j=0}^{\infty} \psi_{j+r}\left(X_{t-j} - \sum_{i=1}^{\infty} a_i X_{t-j-i}\right) \\ &= \sum_{j=0}^{\infty} \phi_j(r) X_{t-j}. \end{aligned} \tag{7.4}$$

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,

$$\begin{aligned} X_t(1) &= \sum_{j=1}^{p} \phi_j X_{t+1-j} \\ X_t(r) &= \sum_{j=r}^{p} \phi_j X_{t+r-j} + \sum_{j=1}^{r-1} \phi_j X_t(r-j) \text{ for } 2 \leq r \leq p \\ X_t(r) &= \sum_{j=1}^{p} \phi_j X_t(r-j) \text{ for } r > p. \end{aligned} \tag{7.5}$$

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*

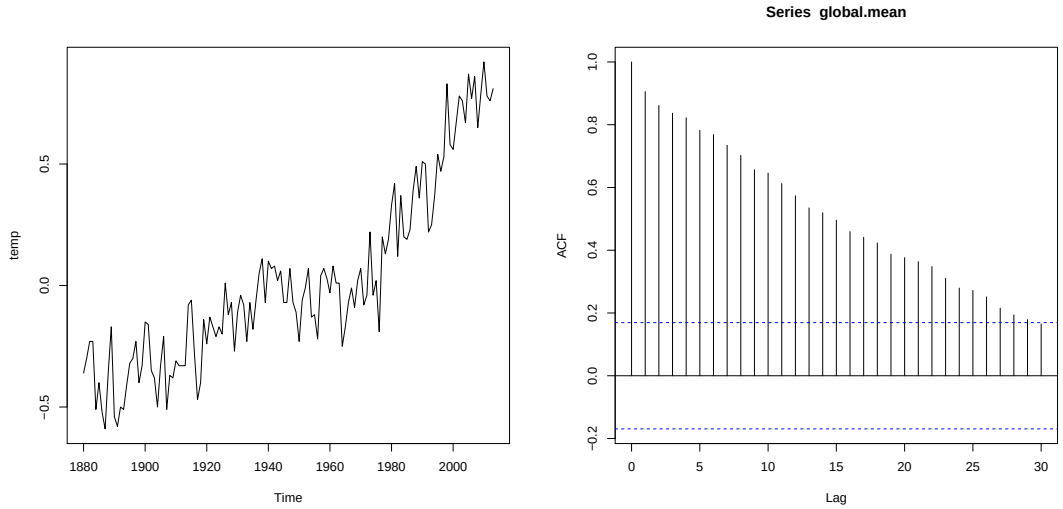


Figure 7.1: Yearly temperature from 1880-2013 and the ACF.

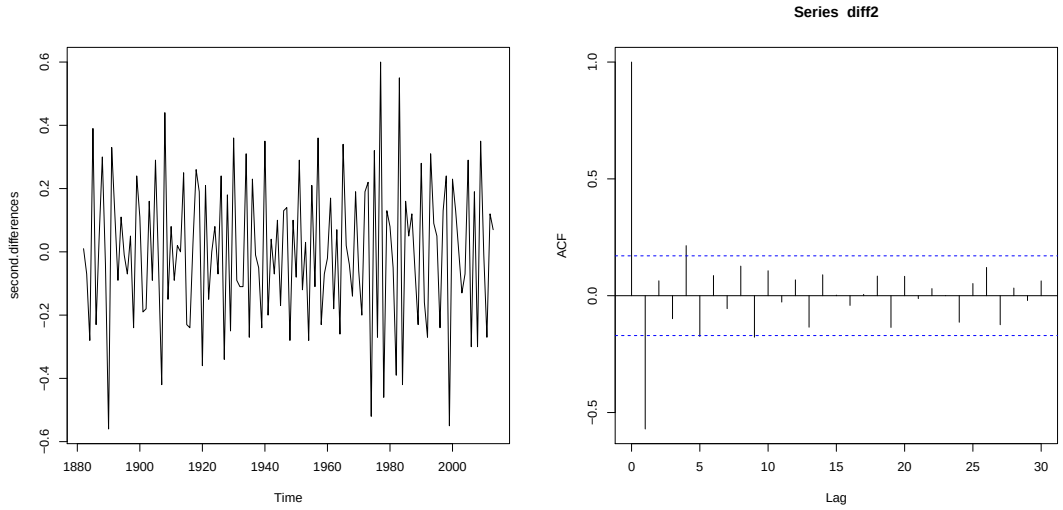


Figure 7.2: Second differences of yearly temperature from 1880-2013 and its ACF.

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

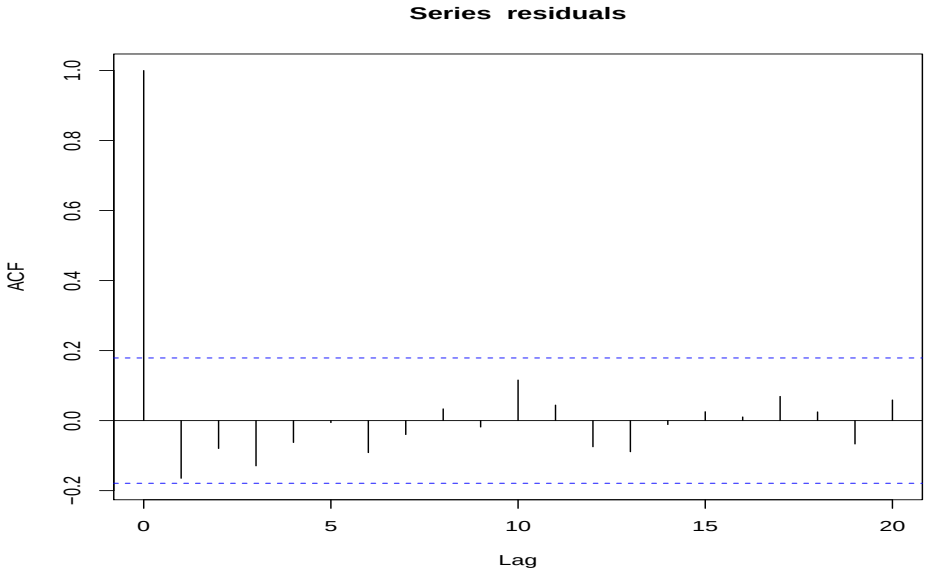


Figure 7.3: An ACF plot of the estimated residuals $\{\widehat{\varepsilon_t}\}$.

$$
\begin{aligned}
\hat{X}_{127}(1) &= -1.1472X_{127} - 1.1565X_{126} - 1.0784X_{125} - 0.7745X_{124} - 0.6132X_{123} \\
&\quad -0.3515X_{122} - 0.1575X_{121} \\
\hat{X}_{127}(2) &= -1.1472\hat{X}_{127}(1) - 1.1565X_{127} - 1.0784X_{126} - 0.7745X_{125} - 0.6132X_{124} \\
&\quad -0.3515X_{123} - 0.1575X_{122} \\
\hat{X}_{127}(3) &= -1.1472\hat{X}_{127}(2) - 1.1565\hat{X}_{127}(1) - 1.0784X_{127} - 0.7745X_{126} - 0.6132X_{125} \\
&\quad -0.3515X_{124} - 0.1575X_{123} \\
\hat{X}_{127}(4) &= -1.1472\hat{X}_{127}(3) - 1.1565\hat{X}_{127}(2) - 1.0784\hat{X}_{127}(1) - 0.7745X_{127} - 0.6132X_{126} \\
&\quad -0.3515X_{125} - 0.1575X_{124} \\
\hat{X}_{127}(5) &= -1.1472\hat{X}_{127}(4) - 1.1565\hat{X}_{127}(3) - 1.0784\hat{X}_{127}(2) - 0.7745\hat{X}_{127}(1) - 0.6132X_{127} \\
&\quad -0.3515X_{126} - 0.1575X_{125}.
\end{aligned}
$$

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

$$\sigma^2 \sum_{j=0}^{k-1} \psi_j(\widehat{\phi})^2 \text{ (using (??))}$$

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.

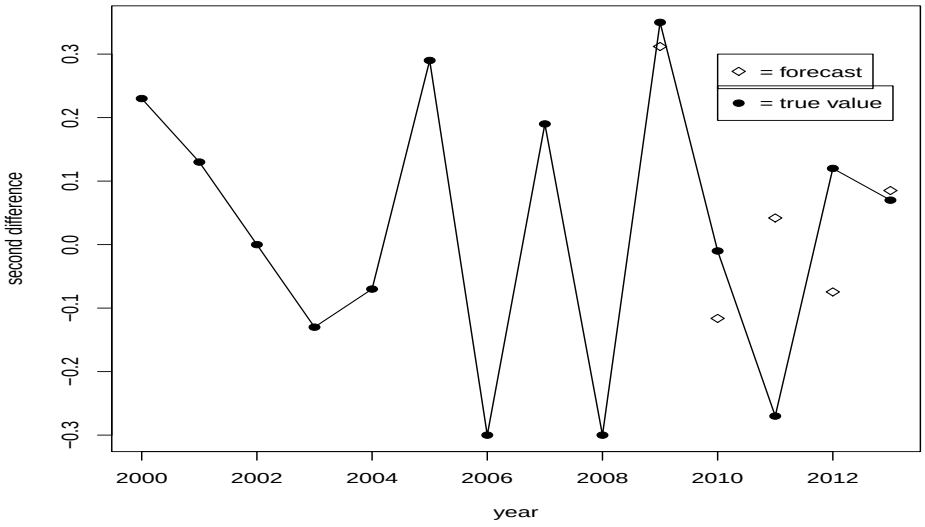


Figure 7.4: Forecasts of second differences.

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

$$X_t = \mu + A\cos(\omega t) + B\sin(\omega t) + \underbrace{\varepsilon_t}_{AR} \quad or$$

$$X_t = \mu + \underbrace{\varepsilon_t}_{AR}$$

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

$$\begin{aligned} X_{t|t-1} &= \sum_{j=1}^{t-1} \phi_{t-1,j} X_{t-j} \qquad P_{X_t,\ldots,X_2}(X_{t+1}) = \sum_{j=1}^{t-1} \phi_{t-1,j} X_{t+1-j} \\ P_{X_t,\ldots,X_2}(X_1) &= \sum_{j=1}^{t-1} \phi_{t-1,j} X_{j+1}. \end{aligned} \tag{7.9}$$

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

$$\begin{aligned} X_{t+1|t} &= P_{X_t,\ldots,X_2}(X_{t+1}) + P_{X_1 - P_{X_t,\ldots,X_2}(X_1)}(X_{t+1}) \\ &= \underbrace{\sum_{j=1}^{t-1} \phi_{t-1,j} X_{t+1-j}}_{\text{by (7.9)}} + \underbrace{\phi_{tt}\left(X_1 - P_{X_t,\ldots,X_2}(X_1)\right)}_{\text{by projection onto one variable}} \\ &= \sum_{j=1}^{t-1} \phi_{t-1,j} X_{t+1-j} + \phi_{t,t} \left(X_1 - \underbrace{\sum_{j=1}^{t-1} \phi_{t-1,j} X_{j+1}}_{\text{by (7.9)}} \right). \end{aligned} \tag{7.10}$$

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

$$L_n = \begin{pmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ -\phi_{1,1} & 1 & 0 & \ldots & \ldots & 0 \\ -\phi_{2,2} & -\phi_{2,1} & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ -\phi_{n-1,n-1} & -\phi_{n-1,n-2} & -\phi_{n-1,n-3} & \ldots & \ldots & 1 \end{pmatrix} \tag{7.14}$$

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

$$\begin{pmatrix} \phi_{p,1}(r) \\ \vdots \\ \phi_{p,p}(r) \end{pmatrix} = \Sigma_p^{-1} \underline{r}_{p,r}, \tag{7.15}$$

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

$$\Phi_p = \begin{pmatrix} \phi_{p,1} & \phi_{p,2} & \phi_3 & \ldots & \phi_{p.p} \\ 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ldots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}.$$

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(∞) repre-

sentation of the ARMA model the best linear predictor is

$$\begin{aligned} X_t(1) &= \sum_{j=1}^{\infty} \psi_j \varepsilon_{t+1-j} \\ &= \sum_{j=1}^{\infty} a_j X_{t+1-j} \end{aligned}$$

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

$$X_t(1) = \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i \left(X_{t+1-i} - X_{t-i}(1) \right).$$

However, for finite predictors the actual one-step ahead prediction formula is not so simple. It can be shown that for $t \geq \max(p, q)$

$$X_{t+1|t} = \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_{t,i} (X_{t+1-i} - X_{t+1-i|t-i}), \tag{7.16}$$

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

$$\widehat{X}_{t+1|t} = \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i (X_{t+1-i} - \widehat{X}_{t+1-i|t-i}). \tag{7.17}$$

This approximation seems plausible, since in the exact predictor (7.16), $\theta_{t,i} \to \theta_i$. By iterating backwards, we can show that

$$\widehat{X}_{t+1|t} = \underbrace{\sum_{j=1}^{t-\max(p,q)} a_j X_{t+1-j}}_{\text{first part of AR}(\infty)\text{ expansion}} + \sum_{j=1}^{\max(p,q)} b_j X_j \tag{7.18}$$

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) 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

$$\widehat{X}_{t+1|t+1} \quad = \quad \widehat{X}_{t+1|t} + K_{t+1}\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right).$$

(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

$$P_{t+1|t+1} \quad = \quad P_{t+1|t} - K_{t+1}HP_{t+1|t} = (I - K_{t+1}H)P_{t+1|t}.$$

(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

$$\begin{pmatrix} X(t+1|t+1) \\ X(t+2|t+1) \\ X(t+3|t) \\ \vdots \\ X(t+r-1|t+1) \\ X(t+r|t+1) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ ? & ? & ? & \dots & ? & ? \end{pmatrix} \begin{pmatrix} X(t|t) \\ X(t+1|t) \\ X(t+2|t) \\ \vdots \\ X(t+r-2|t) \\ X(t+r-1|t) \end{pmatrix} + \varepsilon_{t+1} \begin{pmatrix} 1 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{r-2} \\ \psi_{r-1} \end{pmatrix}.$$

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 $X(t|t), \dots, X(t+r-1|t)$ 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 simplicility 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 $Q = \text{var}(\varepsilon_t)\underline{\psi}'_m\underline{\psi}_m$, $R = 0$, $H = (1, 0, \ldots, 0)$. 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

$$\widehat{Z}_{t+1|t+1} = \widehat{Z}_{t+1|t} + K_{t+1}\left(Y_{t+1} - H\widehat{Z}_{t+1|t}\right).$$

where

$$K_{t+1} = \frac{P_{t+1|t}H^*}{HP_{t+1|t}H^*}$$

and the corresponding mean squared error

$$P_{t+1|t+1} = P_{t+1|t} - K_tHP_{t+1|t} = (I - K_tH)P_{t+1|t}.$$

$\widehat{Z}_{t+1|t}$ 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

$$\begin{aligned} X_{t+1|t}^2 &= \sum_{j=1}^{p} a_j X_{t+1-j}^2 \\ X_{t+k|t}^2 &= \sum_{j=k}^{p} a_j X_{t+k-j}^2 + \sum_{j=1}^{k-1} a_j X_{t+k-j|k}^2 \quad 2 \le k \le p \\ X_{t+k|t}^2 &= \sum_{j=1}^{p} a_j X_{t+k-j|t}^2 \qquad k > p. \end{aligned}$$

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

$$X_t = \phi_1 X_{t-1} + b_{1,1} X_{t-1}\varepsilon_{t-1} + \varepsilon_t.$$

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*

$$X_{t+1} - \widehat{X}_{t+1|t} = -b_{1,1}(X_t - \widehat{X}_{t|t-1})X_t + (b\varepsilon_t X_t + \varepsilon_{t+1}). \tag{7.27}$$

We will compare the above recursion to the recursion based on ε_{t+1}*. Rearranging the bilinear*

equation gives

$$\varepsilon_{t+1} = -b\varepsilon_t X_t + \underbrace{(X_{t+1} - \phi_1 X_t)}_{=b\varepsilon_t X_t + \varepsilon_{t+1}}. \tag{7.28}$$

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 $b^t \prod_{j=1}^{t} \varepsilon_j \stackrel{a.s.}{\to} 0$ *as* $t \to \infty$*, then* $\widehat{X}_{t+1|t} \stackrel{\mathcal{P}}{\to} X_t(1)$ *as* $t \to \infty$*. We now show that if* $\mathrm{E}[\log|\varepsilon_t| < -\log|b|$*, then* $b^t \prod_{j=1}^{t} \varepsilon_j \stackrel{a.s.}{\to} 0$*. 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

$$\log|b^t \prod_{j=1}^{t} \varepsilon_j|^{1/t} = \log|b| + \frac{1}{t}\sum_{j=1}^{t} \log|\varepsilon_j| \stackrel{\mathcal{P}}{\to} \log|b| + \mathrm{E}\log|\varepsilon_0| = \gamma.$$

Thus we see that $|b^t \prod_{j=1}^{t} \varepsilon_j|^{1/t} \stackrel{a.s.}{\to} \exp(\gamma)$*. 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 $MA(\infty)$ process. There is, however, a significant difference. The random variables $\{Z_t\}$ of an $MA(\infty)$ 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

$$\begin{pmatrix} \phi_{1,m} & \phi_{2,m} & \cdots & \phi_{p,m} \\ \phi_{1,m-1} & \phi_{2,m-1} & \cdots & \phi_{p,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{1,m-p+1} & \phi_{2,m-p+1} & \cdots & \phi_{p,m-p+1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} \phi_{1,m-1} & \phi_{2,m-1} & \cdots & \phi_{p,m-1} \\ \phi_{1,m-2} & \phi_{2,m-2} & \cdots & \phi_{p,m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{1,m-p} & \phi_{2,m-p} & \cdots & \phi_{p,m-p} \end{pmatrix}.$$

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

$\{\phi_{\ell,m}; 2 \le \ell \le p\}$ 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

$$\begin{aligned}
\phi_{p,r} &= \phi_p \psi_{r-1} && m - p \le r \le m \\
\phi_{\ell,r} &= \phi_\ell \phi_{1,r-1} + \phi_{\ell+1,r-1} && 1 \le \ell \le p-1, \quad m - p \le r \le m
\end{aligned}$$

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

$$\phi_{p-r,m} = \sum_{s=0}^{r} \phi_{p-r+s}\psi_{m-1-s} \qquad 0 \le r \le p-1.$$

In the last line of the above we change variables with $\ell = p - r$ to give for $m \ge 1$

$$\phi_{\ell,m} = \sum_{s=0}^{p-\ell} \phi_{\ell+s}\psi_{m-1-s} \qquad 1 \le \ell \le p,$$

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

$$\Sigma_t = \begin{pmatrix} c(0) & c(1) & c(2) & \dots & c(t-1) \\ c(1) & c(0) & c(1) & \dots & c(t-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c(t-1) & c(t-2) & \vdots & \vdots & c(0) \end{pmatrix} \quad \text{and} \quad \underline{r}_t = \begin{pmatrix} c(1) \\ c(2) \\ \vdots \\ c(t) \end{pmatrix}.$$

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

$$E_{t-1} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & \vdots & 0 & 0 & 0 \end{pmatrix},$$

(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

$$\underline{\phi}_{t-1,t} = \underbrace{\Sigma_{t-1}^{-1}\underline{r}_{t-1}}_{=\underline{\phi}_{t-1}} - \underbrace{\Sigma_{t-1}^{-1}E_{t-1}\underline{r}_{t-1}}_{=E_{t-1}\underline{\phi}_{t-1}}\phi_{t,t}.$$

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

$$W_{t+1} - \widehat{W}_{t+1|t} = -\theta_1(W_t - \widehat{W}_{t|t-1}) - \theta_2(W_{t-1} - \widehat{W}_{t-1|t-2}) + W_{t+1}. \tag{7.40}$$

It is straightforward to rewrite $W_{t+1} - \widehat{W}_{t+1|t}$ as the matrix difference equation

$$\underbrace{\begin{pmatrix} W_{t+1} - \widehat{W}_{t+1|t} \\ W_t - \widehat{W}_{t|t-1} \end{pmatrix}}_{=\widehat{\underline{\varepsilon}}_{t+1}} = -\underbrace{\begin{pmatrix} \theta_1 & \theta_2 \\ -1 & 0 \end{pmatrix}}_{=Q} \underbrace{\begin{pmatrix} W_t - \widehat{W}_{t|t-1} \\ W_{t-1} - \widehat{W}_{t-1|t-2} \end{pmatrix}}_{=\widehat{\underline{\varepsilon}}_t} + \underbrace{\begin{pmatrix} W_{t+1} \\ 0 \end{pmatrix}}_{\underline{W}_{t+1}}$$

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

$$\underbrace{\begin{pmatrix} \varepsilon_{t+1} \\ \varepsilon_t \end{pmatrix}}_{=\underline{\varepsilon}_{t+1}} = -\underbrace{\begin{pmatrix} \theta_1 & \theta_2 \\ -1 & 0 \end{pmatrix}}_{Q} \underbrace{\begin{pmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{pmatrix}}_{\underline{\varepsilon}_t} + \underbrace{\begin{pmatrix} W_{t+1} \\ 0 \end{pmatrix}}_{\underline{W}_{t+1}} \tag{7.41}$$

Thus iterating backwards we can write

$$\varepsilon_{t+1} = \sum_{j=0}^{\infty} (-1)^j [Q^j]_{(1,1)} W_{t+1-j} = \sum_{j=0}^{\infty} \tilde{b}_j W_{t+1-j},$$

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

$$\begin{aligned} \varepsilon_{t+1} &= \sum_{j=0}^{t-3} (-1)^j [Q^j]_{(1,1)} W_{t+1-j} + (-1)^{t-2} [Q^{t-2}]_{(1,1)} \varepsilon_3 \\ &= \sum_{j=0}^{t-3} \tilde{b}_j W_{t+1-j} + (-1)^{t-2} [Q^{t-2}]_{(1,1)} \varepsilon_3. \end{aligned} \tag{7.42}$$

Therefore, by comparison we see that

$$\varepsilon_{t+1} - \sum_{j=0}^{t-3} \tilde{b}_j W_{t+1-j} = (-1)^{t-2} [Q^{t-2} \underline{\varepsilon}_3]_1 = \sum_{j=t-2}^{\infty} \tilde{b}_j W_{t+1-j}.$$

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, $\text{var}[V_t] = Q$ 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, $\text{var}[W_t] = R$, $\{W_t\}_t$ 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

$$\alpha_X = \frac{\text{cov}(X, Z - P_Y(Z))}{\text{var}(Z - P_Y(Z))}$$

and

$$\text{var}[X - P_{Y,Z}(X)] = \text{cov}[X, X - P_{Y,Z}(X)], \tag{7.44}$$

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 $P_{t+1|t} = \text{var}[X_{t+1} - \widehat{X}_{t+1|t}]$ (predictive) and $\widehat{X}_{t+1|t+1} = P_{Y_1,\ldots,Y_t}(X_{t+1})$ and $P_{t+1|t+1} = \text{var}[X_{t+1} - \widehat{X}_{t+1|t+1}]$ (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

$$\widehat{X}_{t+1|t+1} \;=\; \widehat{X}_{t+1|t} + K_{t+1}\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right).$$

where

$$K_{t+1} = P_{t+1|t}H^*[HP_{t+1|t}H^* + R]^{-1}$$

and the corresponding mean squared error

$$P_{t+1|t+1} \;=\; P_{t+1|t} - K_tHP_{t+1|t} = (I - K_tH)P_{t+1|t}$$

(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:

$$\begin{aligned}\widehat{X}_{t+1|t} = P_{Y_1,\ldots,Y_t}(X_{t+1}) &= P_{Y_1,\ldots,Y_t}(FX_t + V_{t+1}) \\ &= P_{Y_1,\ldots,Y_t}(FX_t) + P_{Y_1,\ldots,Y_t}(V_{t+1}) = FP_{Y_1,\ldots,Y_t}(X_t) = F\widehat{X}_{t|t}.\end{aligned}$$

The mean squared error

$$\begin{aligned} P_{t+1|t} &= \text{var}[X_{t+1} - \widehat{X}_{t+1|t}] = \text{var}[FX_t + V_{t+1} - F\widehat{X}_{t|t}] \\ &= \text{var}[F(X_t - \widehat{X}_{t|t}) + V_{t+1}] \\ &= \text{var}[F(X_t - \widehat{X}_{t|t})] + \text{var}[V_{t+1}] \\ &= F\text{var}[X_t - \widehat{X}_{t|t}]F^* + \text{var}[V_{t+1}] = FP_{t|t}F^* + Q. \end{aligned}$$

This gives the two predictors from the previous update equations. Next the update equations (which is slightly more tricky).

Derivation of the update equations Now we expand the projection space from $\text{sp}(Y_1, \ldots, Y_t)$ to $\text{sp}(Y_1, \ldots, Y_t, Y_{t+1})$. But as the recursion uses $\text{sp}(Y_1, \ldots, Y_t)$ we represent

$$\text{sp}(Y_1, \ldots, Y_t, Y_{t+1}) = \text{sp}(Y_1, \ldots, Y_t, Y_{t+1} - P_{Y_1,\ldots,Y_t}(Y_{t+1})).$$

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

$$\alpha = \text{var}(Y_{t+1} - H\widehat{X}_{t+1|t})^{-1}\text{cov}(X_{t+1}, Y_{t+1} - H\widehat{X}_{t+1|t}).$$

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

$$\begin{aligned} \text{cov}(X_{t+1}, Y_{t+1} - H\widehat{X}_{t+1|t}) &= \text{cov}(X_{t+1}, H(X_{t+1} - X_{t+1|t}) + W_{t+1}) \\ &= \text{cov}(X_{t+1}, H(X_{t+1} - X_{t+1|t})) = \text{cov}(X_{t+1} - X_{t+1|t}, X_{t+1})H^* \\ &= \text{var}(X_{t+1} - X_{t+1|t}) = P_{t+1|t}H^* \end{aligned} \tag{7.45}$$

and

$$\begin{aligned}\text{var}(Y_{t+1} - H\widehat{X}_{t+1|t}) &= \text{var}(H(X_{t+1} - X_{t+1|t}) + W_{t+1}) \\ &= H\text{var}(X_{t+1} - X_{t+1|t})H^* + \text{var}(W_{t+1}) \\ &= HP_{t+1|t}H^* + R.\end{aligned}$$

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

$$\begin{aligned}P_{t+1|t+1} &= \text{var}[X_{t+1} - \widehat{X}_{t+1|t+1}] = \text{cov}[X_{t+1}, X_{t+1} - \widehat{X}_{t+1|t+1}] \\ &= \text{cov}\left[X_{t+1}, X_{t+1} - \widehat{X}_{t+1|t} - K_t\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right)\right] \\ &= \text{cov}\left[X_{t+1}, X_{t+1} - \widehat{X}_{t+1|t}\right] - \text{cov}\left[X_{t+1}, K_t\left(Y_{t+1} - H\widehat{X}_{t+1|t}\right)\right] \\ &= P_{t+1|t} - K_tHP_{t+1|t} = (I - K_tH)P_{t+1|t}\end{aligned}$$

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

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2).$$

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

$$\frac{1}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

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 $S_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

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 ($\sum_k |\text{cov}(X_0, X_k)| < \infty$). 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

$$\text{var}(\bar{Y}_n) = \frac{1}{n} c(0) + \frac{2}{n} \sum_{k=1}^{n} \big(\frac{n-k}{n}\big) c(k).$$

Furthermore, if $\sum_k |c(k)| < \infty$, then in Example 3.3.1 we showed that

$$\text{var}(\bar{Y}_n) \approx \frac{1}{n} c(0) + \frac{2}{n} \sum_{k=1}^{\infty} c(k).$$

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

$$\sqrt{n}(\bar{\epsilon}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1). \tag{8.1}$$

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 $\sqrt{n-q}\bar{\varepsilon}_{n-q} \xrightarrow{\mathcal{P}} \mathcal{N}(0,1)$ as $n \to \infty$. However, for finite q, $\sqrt{(n-q)/n} \xrightarrow{\mathcal{P}} 1$, therefore $\sqrt{n}\bar{\varepsilon}_{n-q} \xrightarrow{\mathcal{P}} \mathcal{N}(0,1)$. Altogether, substituting $\mathrm{E}|E_1| \leq Cn^{-1}$ and $\mathrm{E}|E_2| \leq Cn^{-1}$into (8.2) gives

$$\sqrt{n}\left(\bar{Y}_n - \mu\right) = \Psi\sqrt{n}\bar{\varepsilon}_{n-q} + O_p(\frac{1}{n}) \xrightarrow{\mathcal{P}} \mathcal{N}\left(0, \Psi^2\right).$$

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

$$\mathbf{X}_n = \begin{pmatrix} 0 & 0 & \dots & 0 & X_1 & X_2 & \dots & X_{n-1} & X_n \\ 0 & 0 & \dots & X_1 & X_2 & \dots & X_{n-1} & X_n & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_1 & X_2 & \dots & X_{n-1} & X_n & 0 & \dots & \dots & 0 \end{pmatrix}$$

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 $h \in \{1, \ldots, n\}$

$$\sqrt{n}(\hat{\rho}_n(h) - \rho(h)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, W_h) \tag{8.5}$$

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

$$\sqrt{n}\left(\sum_{j=0}^{d} \alpha_{r_j} \frac{1}{n} \sum_{t=1}^{n-|r_j|} X_t X_{t+r_j} - \sum_{j=0}^{d} \alpha_{r_j} c(r_j) \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_\alpha^2\right), \tag{8.7}$$

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*

$$\sqrt{n}\begin{pmatrix} \widehat{c}_n(0) - c(0) \\ \widehat{c}_n(r_1) - c(r_1) \\ \vdots \\ \widehat{c}_n(r_d) - c(r_d) \end{pmatrix} \xrightarrow{\mathcal{P}} \mathcal{N}(0, V_{d+1})$$

where

$$\begin{aligned}(V_{d+1})_{i,j} &= \sum_{k=-\infty}^{\infty} c(k)c(k + r_{i-1} - r_{j-1}) + \sum_{k=-\infty}^{\infty} c(k + r_{i-1} - 1)c(k - r_{j-1} - 1) + \\ &\quad \sum_{k=-\infty}^{\infty} \kappa_4(r_{i-1} - 1, k, k + r_{j-1} - 1) \end{aligned} \tag{8.8}$$

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*

$$\sqrt{n} \begin{pmatrix} \widehat{\rho}_n(r_1) - \rho(r_1) \\ \vdots \\ \widehat{\rho}_n(r_d) - \rho(r_d) \end{pmatrix} \xrightarrow{\mathcal{P}} \mathcal{N}\left(0, G(C_{d+1} + K_{d+1})G'\right)$$

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

$$G = \frac{1}{c(0)} \begin{pmatrix} -\rho(r_1) & 1 & 0 & \dots & 0 \\ -\rho(r_2) & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \ddots & 0 \\ -\rho(r_d) & 0 & \dots & \dots & 1 \end{pmatrix}$$

PROOF. We define the $g : \mathbb{R}^{d+1} \to \mathbb{R}^d$ vector function

$$g(x_0, x_1, \dots, x_d) = \left(\frac{x_1}{x_0}, \dots, \frac{x_d}{x_0}\right).$$

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

$$\mathcal{N}(G) = \{\alpha \underline{c}_{d+1}; \alpha \in \mathbb{R}\}$$

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*

$$\sqrt{n}\begin{pmatrix} \widehat{\rho}_n(r_1) - \rho(r_1) \\ \vdots \\ \widehat{\rho}_n(r_d) - \rho(r_d) \end{pmatrix} \stackrel{\mathcal{P}}{\to} \mathcal{N}\left(0, GC_{d+1}G'\right),$$

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

$$\sqrt{n}\begin{pmatrix} \widehat{\rho}_n(1) \\ \vdots \\ \widehat{\rho}_n(h) \end{pmatrix} \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, I_h).$$

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 $\text{cum}(X_{t+k_i}, X_{t+k_j})$ (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

$$\text{cum}\,(X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}) = 0.$$

  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

$$\begin{aligned} & \text{cum}\,(aX_1 + bY_1 + c, X_2, X_3, X_4) \\ = \;& a\text{cum}\,(X_1, X_2, X_3, X_4) + b\text{cum}\,(Y_1, X_2, X_3, X_4) \end{aligned}$$

  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

$$\text{cum}(X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}) = \text{cum}(X_0, X_{k_1}, X_{k_2}, X_{k_3}) = \kappa_4(k_1, k_2, k_3).$$

Comparisons between the covariance and higher order cumulants

(a) The covariance is invariant to ordering $\text{cov}[X_t, X_{t+k}] = \text{cov}[X_{t+k}, X_t]$.

Like the covariance, the joint cumulant $\text{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}]$ is also invariant to order.

(b) The covariance $\text{cov}[X_t, X_{t+k}]$ is a measure of linear dependence between X_t and X_{t+k}.

The cumulant is measuring the dependence between $\text{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}]$ 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

$$\begin{aligned} & \text{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}] \\ = \; & \text{E}[X_t X_{t+k_1} X_{t+k_2} X_{t+k_3}] - \text{E}[X_t X_{t+k_1}]\text{E}[X_{t+k_2} X_{t+k_3}] \\ & -\text{E}[X_t X_{t+k_2}]\text{E}[X_{t+k_1} X_{t+k_3}] - \text{E}[X_t X_{t+k_3}]\text{E}[X_{t+k_1} X_{t+k_2}]. \end{aligned} \tag{8.12}$$

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 $k > 0$

$$|\text{cov}[X_t, X_{t+k}]| \leq \alpha(k)$$

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 $\text{cov}(X_t, X_{t+h}) = \sigma^2 \sum_j \psi_j \psi_{j+h}$. 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

$$|\text{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}]| \leq \alpha(k_1)\alpha(k_2-k_1)\alpha(k_3-k_2) \tag{8.13}$$

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 $\kappa_4 = \text{cum}(\varepsilon_t, \varepsilon_t, \varepsilon_t, \varepsilon_t)$). 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*

$$\begin{aligned} cum[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}] &= \sum_{j_0,j_1,j_2,j_3=0}^{\infty} \phi^{j_0+j_1+j_2+j_3} \text{cum}\left[\varepsilon_{t-j_0}, \varepsilon_{t+k_1-j_1}, \varepsilon_{t+k_2-j_2}, \varepsilon_{t+k_3-j_3}\right] \\ &= \kappa_4 \sum_{j=0}^{\infty} \phi^j \phi^{j+k_1} \phi^{j+k_2} \phi^{j+k_3} = \kappa_4 \frac{\phi^{k_1+k_2+k_3}}{1-\phi^4}. \end{aligned}$$

The fourth order dependence decays as the lag increases. And this rate of decay is faster than the general bound $|\text{cum}[X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}]| \leq \alpha(k_1)\alpha(k_2-k_1)\alpha(k_3-k_2)$.

- *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

$$\begin{aligned}(V_{d+1})_{i,j} &= \sum_{k=-\infty}^{\infty} c(k)c(k+r_{i-1}-r_{j-1}) + \sum_{k=-\infty}^{\infty} c(k+r_{i-1}-1)c(k-r_{j-1}-1) + \\ &\quad \sum_{k=-\infty}^{\infty} \kappa_4(r_{i-1}-1,k,k+r_{j-1}-1). \end{aligned} \tag{8.14}$$

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

$$\begin{pmatrix} c(0)^2 & c(1)^2 & \dots & c(n-1)^2 \\ c(-1)^2 & c(0)^2 & \dots & c(n-2)^2 \\ \vdots & \vdots & \ddots & \vdots \\ c((n-1))^2 & c((n-2))^2 & \dots & c(0)^2 \end{pmatrix},$$

(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

$$\sqrt{n}\begin{pmatrix} \widehat{\rho}_n(r_1) - \rho(r_1) \\ \vdots \\ \widehat{\rho}_n(r_d) - \rho(r_d) \end{pmatrix} \stackrel{\mathcal{P}}{\to} \mathcal{N}\left(0, GC_{d+1}G'\right).$$

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

$$\sum_{k=-\infty}^{\infty} \mathrm{cum}(X_0, X_{r_1}, X_k, X_{k+r_2}) = \kappa_4 \sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j-r_1} \psi_{j-k} \psi_{j-r_2-k}.$$

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

$$(K_{d+1})_{i,j} = \sum_{k=-\infty}^{\infty} \kappa_4(r_{i-1}, k, k + r_{j-1}) = \frac{\kappa_4}{\sigma^2} c(r_{i-1}) c(j_{i-1})$$

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, $\sqrt{n}\hat{\boldsymbol{\rho}}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_h)$. 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

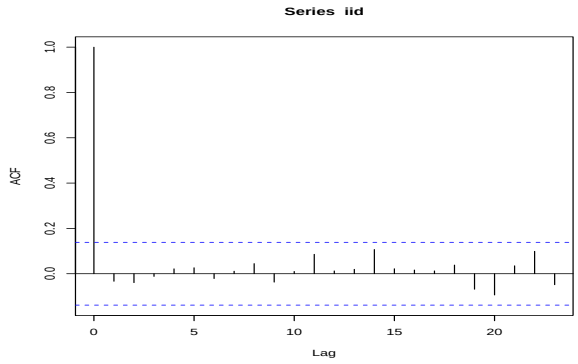


Figure 8.1: The sample ACF of an iid sample with error bars (sample size $n = 200$).

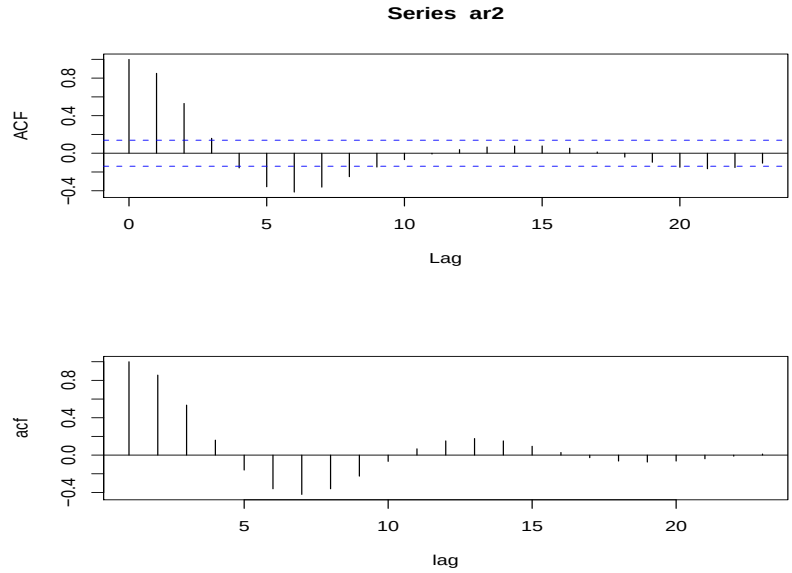


Figure 8.2: Top: The sample ACF of the AR(2) process $X_t = 1.5X_{t-1} + 0.75X_{t-2} + \varepsilon_t$ with error bars $n = 200$. Bottom: The true ACF.

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 8.5.

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

$$\text{cov}\left(\sqrt{n}\widehat{c}_n(r_1), \sqrt{n}\widehat{c}_n(r_2)\right) = \begin{cases} \sum_k \kappa_4(r_1, k, k+r_2) & r_1 \neq r_2 \\ c(0)^2 + \sum_k \kappa_4(r, k, k+r) & r_1 = r_2 = (r) \end{cases}$$

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

$$\text{E}(X_t|X_{t-1}, X_{t-2}, X_{t-3}, \ldots) = 0.$$

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

$$\begin{aligned} \text{cov}(X_t, X_{t+r}) &= \text{E}(X_tX_{t+r}) - \text{E}(X_t)\text{E}(X_{t+r}) \\ &= \text{E}(X_t\text{E}(X_{t+r}|X_t)) - \text{E}(\text{E}(X_t|X_{t-1}))\text{E}(\text{E}(X_{t+r}|X_{t+r-1})) = 0. \end{aligned}$$

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:

$$\mathrm{E}[X_t^2 X_{s_1} X_{s_2}] = 0 \qquad t > s_1, s_2. \tag{8.17}$$

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

$$\mathrm{E}(X_t X_{t+r_1} X_t X_{t+r_2}) = \mathrm{E}\left(X_t^2 X_{t+r_1}\mathrm{E}(X_{t+r_2}|\underbrace{X_t^2, X_{t+r_1}}_{\subset\sigma(X_{t+r_2-1}, X_{t+r_2-2}, \ldots)})\right) = 0.$$

However, in the case that $t_1 + r_1 = t_2 + r_2$ $(r_1, r_2 \geq 0)$ we have

$$\mathrm{E}(X_{t_1+r_1}^2 X_{t_1} X_t X_{t_2}) \neq 0,$$

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 $\mathrm{E}[X_t^2 X_{s_1} X_{s_2}] = 0$ for $t > s_1, s_2$, 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

$$\sqrt{n}\begin{pmatrix} \widehat{c}_n(1)/\sigma_1 \\ \widehat{c}_n(2)/\sigma_2 \\ \vdots \\ \widehat{c}_n(h)/\sigma_h \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I_h\right).$$

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 $\mathcal{R}_h \xrightarrow{\mathcal{D}} \chi_h^2$ (for h kept fixed). 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.

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

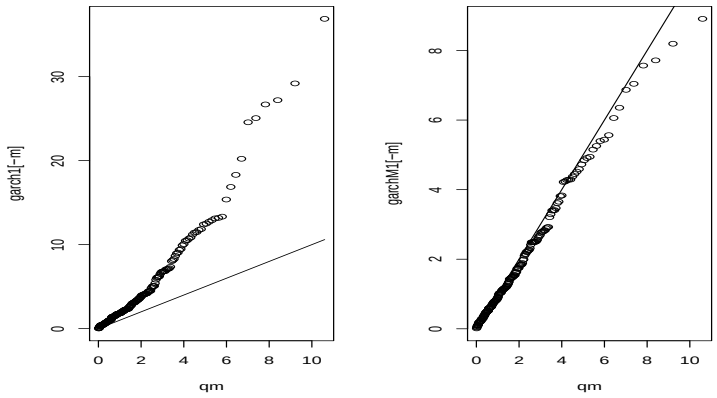


Figure 8.3: Using ARCH(1) time series over 200 replications Left: $\mathcal{S}_2$ against the quantiles of a chi-square distribution with 2df for an ARCH process. Right: $\mathcal{R}_2$ against the quantiles of a chi-square distribution with 2df for an ARCH process.

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.

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 and Philips (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

$$\sqrt{n}\left(\widehat{\boldsymbol{\phi}}_m - \boldsymbol{\phi}_m\right) \stackrel{\mathcal{P}}{\to} N(0, \sigma^2 \Sigma_m^{-1}).$$

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

$$\sqrt{n}\widehat{\phi}_{mm} \stackrel{\mathcal{P}}{\to} N(0, \sigma^2 \Sigma^{mm}).$$

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

$$\sqrt{n}\widehat{\phi}_{mm} \stackrel{\mathcal{P}}{\to} N(0, 1).$$

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 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.

$$\widehat{\theta} = \arg\min \sum_{t=1}^{n} (X_t - g(Y_t, \theta))^2 . \tag{8.19}$$

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

$$\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} = -2 \sum_{t=1}^{n} [X_t - g(Y_t, \theta)] \frac{\partial g(Y_t, \theta)}{\partial \theta}$$

and

$$\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2} = -2\sum_{t=1}^{n} [X_t - g(Y_t,\theta)] \frac{\partial^2 g(Y_t,\theta)}{\partial \theta^2} + 2\sum_{t=1}^{n} \left(\frac{\partial g(Y_t,\theta)}{\partial \theta}\right)^2,$$

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 (8.19) 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 $c(0) = \text{var}[X_t]$, $\sigma^2 = \text{var}(\varepsilon_t)$ and $\mathcal{J}_r = \text{E}[\frac{\partial g(Y_{t+r},\theta)}{\partial \theta}\varepsilon_t]$ and $\mathcal{I} = \text{E}\left(\frac{\partial g(Y_t,\theta)}{\partial \theta}\right)^2$ (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 8.5.1 *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

$$\begin{aligned}\sqrt{n}\widehat{c}_n(r) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n-r}\left(\varepsilon_t - (\widehat{\theta}-\theta)\frac{\partial g(\theta,Y_t)}{\partial\theta} + \frac{(\widehat{\theta}-\theta)^2}{2}\frac{\partial^2 g(\bar{\theta}_t,Y_t)}{\partial\theta^2}\right)\times \\ &\quad \left(\varepsilon_{t+r} - (\widehat{\theta}-\theta)\frac{\partial g(\theta,Y_{t+r})}{\partial\theta} + \frac{(\widehat{\theta}-\theta)^2}{2}\frac{\partial^2 g(\bar{\theta}_{t+r},Y_{t+r})}{\partial\theta^2}\right) \\ &= \sqrt{n}\widetilde{c}_n(r) - \sqrt{n}(\widehat{\theta}-\theta)\frac{1}{n}\sum_{t=1}^{n-r}\left(\varepsilon_t\frac{\partial g(\theta,Y_{t+r})}{\partial\theta} + \varepsilon_{t+r}\frac{\partial g(\theta,Y_t)}{\partial\theta}\right) + O_p(n^{-1/2}) \quad (8.20)\end{aligned}$$

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

$$\frac{1}{n}\sum_{t=1}^{n-r}\left(\varepsilon_t\frac{\partial g(\theta,Y_{t+r})}{\partial\theta} + \varepsilon_{t+r}\frac{\partial g(\theta,Y_t)}{\partial\theta}\right) \stackrel{\text{a.s.}}{\to} \mathrm{E}\left(\varepsilon_t\frac{\partial g(\theta,Y_{t+r})}{\partial\theta}\right) = \mathcal{J}_r,$$

where we use that ε_{t+r} and $\frac{\partial g(\theta,Y_t)}{\partial\theta}$ are independent. Subsituting this into (8.20) 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);

$$\begin{aligned}&\mathrm{var}\left[\sqrt{n}\widetilde{c}_n(r) + \frac{\sqrt{n}}{2}\mathcal{I}^{-1}\mathcal{J}_r\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\right] \\ &\underbrace{\mathrm{var}\left(\sqrt{n}\widetilde{c}_n(r)\right)}_{=1} + 2\mathcal{I}^{-1}\mathcal{J}_r\mathrm{cov}\left(\sqrt{n}\widetilde{c}_n(r), \frac{\sqrt{n}}{2}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\right) + \mathcal{I}^{-2}\mathcal{J}_r^2\mathrm{var}\left(\frac{\sqrt{n}}{2}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\right) \quad (8.21)\end{aligned}$$

We evaluate the two covariance above;

$$\begin{aligned}
&\operatorname{cov}\left(\sqrt{n}\widetilde{c}_n(r), -\frac{\sqrt{n}}{2}\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\right) = \frac{1}{n}\sum_{t_1,t_2=1}^{n-r}\left[\operatorname{cov}\left\{\varepsilon_{t_1}\varepsilon_{t_1+r}, \varepsilon_{t_2}\frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right\}\right] \\
= \; &\frac{1}{n}\sum_{t_1,t_2=1}^{n-r}\left[\operatorname{cov}\{\varepsilon_{t_1},\varepsilon_{t_2}\}\operatorname{cov}\left\{\varepsilon_{t_1+r}, \frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right\} + \operatorname{cov}\{\varepsilon_{t_1+r},\varepsilon_{t_2}\}\operatorname{cov}\left\{\varepsilon_{t_1}, \frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right\}\right. \\
&\left.+\operatorname{cum}\left\{\varepsilon_{t_1},\varepsilon_{t_1+r},\varepsilon_{t_2},\frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right\}\right] = \sigma^2\mathrm{E}\left[\varepsilon_t\frac{\partial g(Y_{t+r},\theta)}{\partial\theta}\right] = \sigma^2\mathcal{J}_r.
\end{aligned}$$

Similarly we have

$$\operatorname{var}\left(\frac{\sqrt{n}}{2}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\right) = \frac{1}{n}\sum_{t_1,t_2=1}^{n}\operatorname{cov}\left(\varepsilon_{t_1}\frac{\partial g(Y_{t_1},\theta)}{\partial\theta}, \varepsilon_{t_2}\frac{\partial g(Y_{t_2},\theta)}{\partial\theta}\right) = \sigma^2\mathrm{E}\left(\frac{\partial g(Y_{t_1},\theta)}{\partial\theta}\right)^2 = \sigma^2\mathcal{I}.$$

Substituting the above into (8.21) 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

$$\sqrt{n}\widehat{\rho}(r) = \mathcal{N}\left(0, 1 - \frac{\sigma^2}{c(0)}\mathcal{J}_r\mathcal{I}^{-1}\mathcal{J}_r\right).$$

8.6 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?

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

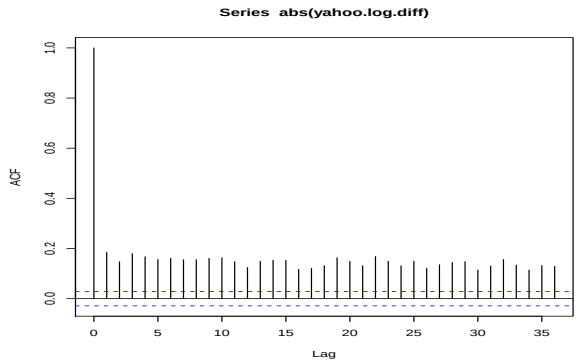


Figure 8.4: ACF plot of the absolute of the log differences.

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 $\hat{r}_n(2) = \hat{c}_n(0) - \hat{\phi}_{1,1}\hat{c}_n(1)$.

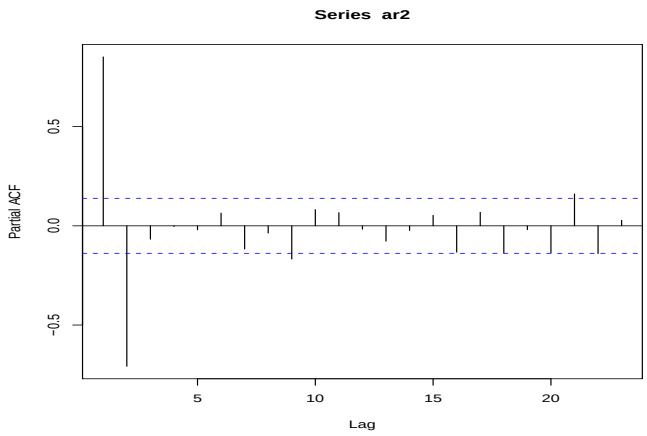


Figure 9.1: Top: The sample partial autocorrelation plot of the AR(2) process $X_t = 1.5X_{t-1} + 0.75X_{t-2} + \varepsilon_t$ with error bars $n = 200$.

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. □

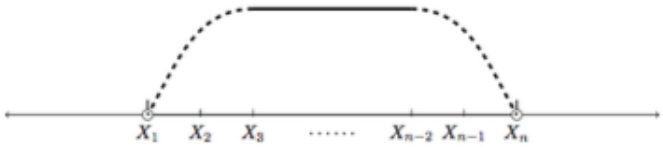
From the above result we can immediately see that the Yule-Walker estimators of the AR(p) coefficients yield a causal solution. Since the autocovariance estimators $\{\widehat{c}_n(r)\}$ form a positive semi-definite sequence, there exists a vector $\underline{Y}_p$ where $\mathrm{var}_{\widehat{\Sigma}_{p+1}}[\underline{Y}_{p+1}] = \widehat{\Sigma}_{p+1}$ with $(\widehat{\Sigma}_{p+1}) = \widehat{c}_n(i-j)$, thus by the above lemma we have that $\widehat{\Sigma}_p^{-1}\widehat{r}_p$ 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*

$$\begin{aligned} \mathbb{P}(X_{t+1} \leq x_{t+1} | X_t = x_t, \ldots, X_1 = x_1) &= \mathbb{P}(\underbrace{\phi X_t + \varepsilon_t \leq x_{t+1}}_{\text{all information contained in } X_t} | X_t = x_t, \ldots, X_1 = x_1) \\ &= \mathbb{P}_\varepsilon(\varepsilon_t \leq x_{t+1} - \phi x_t) = \mathbb{P}(X_{t+1} \leq x_{t+1} | X_t = x_t), \end{aligned}$$

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

$$\begin{aligned}\mathcal{L}_n(\underline{a}) &= \log|\Sigma_n(\underline{a})| + \underline{X}_p'\Sigma_n(\underline{a})^{-1}\underline{X}_p \\ &= \log|\Sigma_p(\underline{a})| + \underline{X}_p'\Sigma_p(\underline{a})^{-1}\underline{X}_p + \underbrace{L_n(\underline{a};\underline{X})}_{\text{conditional likelihood}} . \end{aligned} \tag{9.9}$$

The maximum likelihood estimator is

$$\widehat{\underline{\phi}}_n = \arg\max_{\underline{a}\in\Theta} \mathcal{L}_n(\underline{a}). \tag{9.10}$$

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

$$L_n(\underline{a};\underline{X}) = -(n-p)\log\sigma^2 - \frac{1}{\sigma^2}\sum_{t=p}^{n-1}\left(X_{t+1} - \sum_{j=1}^{p} a_j X_{t+1-j}\right)^2,$$

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

$$\begin{aligned}\widetilde{\underline{\phi}}_p &= \arg\min \sum_{t=p}^{n-1}\left(X_{t+1} - \sum_{j=1}^{p} a_j X_{t+1-j}\right)^2 \\ &= \widetilde{\Sigma}_p^{-1}\widetilde{\underline{r}}_p,\end{aligned}$$

where $(\widetilde{\Sigma}_p)_{i,j} = \frac{1}{n-p}\sum_{t=p+1}^{n} X_{t-i}X_{t-j}$ and $(\widetilde{r}_n)_i = \frac{1}{n-p}\sum_{t=p+1}^{n} X_t X_{t-i}$.

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*

$$\widehat{\phi}_{LS} = \frac{X_1 X_2}{X_1^2}$$

whereas the Yule-Walker estimator of the $AR(1)$ *parameter is*

$$\widehat{\phi}_{YW} = \frac{X_1 X_2}{X_1^2 + X_2^2}.$$

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. □

9.1.6 Sampling properties of the AR regressive estimators

Both the Yule-Walker, least squares and Gaussian likelihood estimator have the same asymptotic sampling properties (under the assumption of stationarity) (the estimator for the tapered Yule-Walker is a little different). Suppose that $\{X_t\}$ has a causal AR(p_0) representation

$$X_t = \sum_{j=1}^{p_0} \phi_j X_{t-j} + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid random variables with $\text{var}[\varepsilon_t] = \sigma^2$ and $\text{E}[|\varepsilon_t|^{2+\delta}] < \infty$ for some $\delta > 0$.

Suppose the AR(p) model is fitted to the time series, using either least squares or Yule-Walker estimator. We denote this estimator as $\widehat{\phi}_p$. If $p \geq p_0$, then

$$\sqrt{n}(\widehat{\underline{\phi}}_p - \underline{\phi}_p) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2 \Sigma_p^{-1}\right),$$

where $\Sigma_p = \text{var}[\underline{X}_p]$ and $\underline{X}_p = (X_1, \ldots, X_p)$.

Remark 9.1.4 *We note that the assumption* $\text{E}|\varepsilon_t^{2+\delta}| < \infty$ *implies that* $\text{E}[|X_t|^{2+\delta}] < \infty$. *In the proof below we use the stronger assumption* $\text{E}(\varepsilon_t^4) < \infty$ *to make the proof easier to follow.*

Tools to prove the result: Martingale central limit theorem (advanced)

We summarize the result, see Billingsley (1995) Hall and Heyde (1980) (Theorem 3.2 and Corollary 3.1) for the details.

Definition 9.1.1 *The random variables* $\{Z_t\}$ *are called martingale differences if*

$$\text{E}(Z_t | Z_{t-1}, Z_{t-2}, \ldots) = 0.$$

The sequence $\{\mathcal{S}_n\}_n$, *where*

$$\mathcal{S}_n = \sum_{t=1}^{n} Z_t$$

are called martingales if $\{Z_t\}$ *are martingale differences. Observe that* $\text{E}[\mathcal{S}_n | \mathcal{S}_{n-1}] = \mathcal{S}_{n-1}$.

Remark 9.1.5 (Martingales and covariances) *We observe that if* $\{Z_t\}$ *are martingale differ-*

ences then

$$\mathrm{E}[Z_t] = \mathrm{E}[\mathrm{E}[Z_t|\mathcal{F}_{t-1}]] = 0,$$

where $\mathcal{F}_s = \sigma(Z_s, Z_{s-1}, \ldots)$ *and for* $t > s$ *and*

$$\mathrm{cov}(Z_s, Z_t) = \mathrm{E}(Z_s Z_t) = \mathrm{E}\big(\mathrm{E}(Z_s Z_t|\mathcal{F}_s)\big) = \mathrm{E}\big(Z_s \mathrm{E}(Z_t|\mathcal{F}_s)\big) = \mathrm{E}(Z_s \times 0) = 0.$$

Hence martingale differences are uncorrelated.

Example 9.1.3 *Suppose that* $X_t = \phi X_{t-1} + \varepsilon_t$*, where* $\{\varepsilon_t\}$ *are iid r.v. with* $\mathrm{E}(\varepsilon_t) = 0$ *and* $|\phi| < 1$*. Then* $\{\varepsilon_t X_{t-1}\}_t$ *are martingale differences. To see why note that*

$$\begin{aligned} \mathrm{E}\left[\varepsilon_t X_{t-1} | \varepsilon_{t-j} X_{t-j-1}; j \geq 1\right] &= \mathrm{E}\left[\mathrm{E}\left(\varepsilon_t X_{t-1} | \varepsilon_{t-j}; j \geq 1\right) | \varepsilon_{t-j} X_{t-j-1}; j \geq 1\right] \\ &= \mathrm{E}\left[X_{t-1} \mathrm{E}\left(\varepsilon_t | \varepsilon_{t-j}; j \geq 1\right) | \varepsilon_{t-j} X_{t-j-1}; j \geq 1\right] = 0, a.s \end{aligned}$$

since $\sigma(\varepsilon_{t-j} X_{t-j-1}; j \geq 1) \subseteq \sigma(\varepsilon_{t-j}; j \geq 1)$*. In general, if* X_t *is a causal time series then* $\{\varepsilon_t X_{t-j}\}_t$ *are martingale differences* $(j > 0)$*.*

Let

$$S_n = \frac{1}{n} \sum_{t=1}^{n} Z_t, \tag{9.11}$$

and $\mathcal{F}_t = \sigma(Z_t, Z_{t-1}, \ldots)$, $\mathrm{E}(Z_t|\mathcal{F}_{t-1}) = 0$ and $\mathrm{E}(Z_t^2) < \infty$. We shall show asymptotic normality of $\sqrt{n}(S_n - \mathrm{E}(S_n))$. The reason for normalising by $\sqrt{n}$, is that $(S_n - \mathrm{E}(S_n)) \xrightarrow{\mathcal{P}} 0$ as $n \to \infty$, hence in terms of distributions it converges towards the point mass at zero. Therefore we need to increase the magnitude of the difference. If it can show that $\mathrm{var}(S_n) = O(n^{-1})$, then $\sqrt{n}(S_n - \mathrm{E}(\mathcal{S}_0) = O(1)$.

Theorem 9.1.1 *Let* S_n *be defined as in (14.16). Further suppose*

$$\frac{1}{n} \sum_{t=1}^{n} Z_t^2 \xrightarrow{\mathcal{P}} \sigma^2, \tag{9.12}$$

where σ^2 *is a finite constant, for all* $\eta > 0$*,*

$$\frac{1}{n} \sum_{t=1}^{n} \mathrm{E}(Z_t^2 I(|Z_t| > \eta\sqrt{n}) | \mathcal{F}_{t-1}) \xrightarrow{\mathcal{P}} 0, \tag{9.13}$$

(this is known as the conditional Lindeberg condition) and

$$\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}(Z_t^2|\mathcal{F}_{t-1}) \xrightarrow{\mathcal{P}} \sigma^2. \tag{9.14}$$

Then we have

$$n^{1/2}S_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2). \tag{9.15}$$

Remark 9.1.6 (The conditional likelihood and martingales) *It is interesting to note that the derivative of conditional log-likelihood of a time series at the true parameter* **is** *a martingale so long as the likelihood is correctly specified. In other works, using that*

$$\log f(X_n,\ldots,X_1|X_1;\theta) \;=\; \sum_{t=2}^{n}\log f(X_t|X_{t-1},\ldots,X_1;\theta),$$

then $\frac{\partial \log f(X_t|X_{t-1},\ldots,X_1;\theta)}{\partial\theta}$ *is a martingale difference. To see why, note that if we can take the derivative outside the integral then*

$$\begin{aligned}
&\mathrm{E}\left[\frac{\partial \log f(X_t|X_{t-1},\ldots,X_1;\theta)}{\partial\theta}\Big|X_{t-1},\ldots,X_1\right]\\
=&\int\frac{\partial \log f(X_t|X_{t-1},\ldots,X_1;\theta)}{\partial\theta}f(X_t|X_{t-1},\ldots,X_1;\theta)dX_t\\
=&\int\frac{\partial f(X_t|X_{t-1},\ldots,X_1;\theta)}{\partial\theta}dX_t = \frac{\partial}{\partial\theta}\int f(X_t|X_{t-1},\ldots,X_1;\theta)dX_t = 0.
\end{aligned}$$

Asymptotic normality of the least squares estimator of the AR(1) parameter

In this section we show asymptotic normality of the least squares estimator of the AR(1), where $\hat{\phi}_n = \arg\max \mathcal{L}_n(a)$ and

$$\mathcal{L}_n(a) = \frac{1}{n-1}\sum_{t=2}^{n}(X_t - aX_{t-1})^2.$$

The first and the second derivative (at the true parameter) is

$$\begin{aligned}\nabla\mathcal{L}_n(a)\rfloor_{a=\phi} &= \frac{-2}{n-1}\sum_{t=2}^{n} X_{t-1}\underbrace{(X_t-\phi X_{t-1})}_{=\varepsilon_t} = \frac{-2}{n-1}\sum_{t=2}^{n} X_{t-1}\varepsilon_t \\ \text{and } \nabla^2\mathcal{L}_n(a) &= \frac{2}{n-1}\sum_{t=2}^{n} X_{t-1}^2 \text{ (does not depend on unknown parameters).}\end{aligned}$$

Thus it is clear that

$$(\widehat{\phi}_n-\phi) = -\left(\nabla^2\mathcal{L}_n\right)^{-1}\nabla\mathcal{L}_n(\phi). \tag{9.16}$$

Since $\{X_t^2\}$ are ergodic random variables, by using the ergodic theorem we have $\nabla^2\mathcal{L}_n \stackrel{\text{a.s.}}{\to} 2\mathrm{E}(X_0^2)$. This, together with (9.16), implies

$$\begin{aligned}\sqrt{n}(\widehat{\phi}_n-\phi) &= \frac{\sum_{t=2}^n X_tX_{t-1}}{\sum_{t=2}^n X_{t-1}^2}-\phi \\ &= \frac{\sum_{t=2}^n X_{t-1}(X_t-\phi X_{t-1})}{\sum_{t=2}^n X_{t-1}^2} = \frac{\sum_{t=2}^n X_{t-1}\varepsilon_t}{\sum_{t=2}^n X_{t-1}^2} \\ &= -\underbrace{\left(\nabla^2\mathcal{L}_n\right)^{-1}}_{\stackrel{\text{a.s.}}{\to}(2\mathrm{E}(X_0^2))^{-1}}\sqrt{n}\nabla\mathcal{L}_n(\phi) = -\Sigma_1^{-1}\sqrt{n}S_n + O_p(n^{-1/2}),\end{aligned}$$

where $S_n = \frac{1}{n-1}\sum_{t=2}^n X_{t-1}\varepsilon_t$. Thus to show asymptotic normality of $\sqrt{n}(\widehat{\phi}_n-\phi)$, will need only show asymptotic normality of $\sqrt{n}S_n$. S_n is the sum of martingale differences, since $\mathrm{E}(X_{t-1}\varepsilon_t|X_{t-1}) = X_{t-1}\mathrm{E}(\varepsilon_t|X_{t-1}) = X_{t-1}\mathrm{E}(\varepsilon_t) = 0$, therefore we apply the martingale central limit theorem (summarized in the previous section).

To show that $\sqrt{n}S_n$ is asymptotically normal, we need to verify conditions (9.12)-(9.14). We note in our example that $Z_t := X_{t-1}\varepsilon_t$, and that the series $\{X_{t-1}\varepsilon_t\}_t$ is an ergodic process (this simply means that sample means converge almost surely to their expectation, so it is a great tool to use). Furthermore, since for any function g, $\mathrm{E}(g(X_{t-1}\varepsilon_t)|\mathcal{F}_{t-1}) = \mathrm{E}(g(X_{t-1}\varepsilon_t)|X_{t-1})$, where $\mathcal{F}_{t-1} = \sigma(X_{t-1}, X_{t-2}, \ldots)$ we need only to condition on X_{t-1} rather than the entire sigma-algebra $\mathcal{F}_{t-1}$. To simplify the notation we let $S_n = \frac{1}{n}\sum_{t=1}^n \varepsilon_t X_{t-1}$ (included an extra term here).

Verification of conditions

C1 : By using the ergodicity of $\{X_{t-1}\varepsilon_t\}_t$ we have

$$\frac{1}{n}\sum_{t=1}^{n} Z_t^2 = \frac{1}{n}\sum_{t=1}^{n} X_{t-1}^2\varepsilon_t^2 \stackrel{\mathcal{P}}{\to} \mathrm{E}(X_{t-1}^2)\underbrace{\mathrm{E}(\varepsilon_t^2)}_{=1} = \sigma^2 c(0).$$

C2 : We now verify the conditional Lindeberg condition.

$$\frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(Z_t^2 I(|Z_t| > \eta\sqrt{n})|\mathcal{F}_{t-1}) = \frac{1}{n-1}\sum_{t=1}^{n} \mathrm{E}(X_{t-1}^2\varepsilon_t^2 I(|X_{t-1}\varepsilon_t| > \eta\sqrt{n})|X_{t-1}).$$

We now use the Cauchy-Schwartz inequality for conditional expectations to split $X_{t-1}^2\varepsilon_t^2$ and $I(|X_{t-1}\varepsilon_t| > \varepsilon)$ (see the conditional Hölder inequality). We recall that the Cauchy-Schwartz inequality for conditional expectations is $\mathrm{E}(X_tZ_t|\mathcal{G}) \leq [\mathrm{E}(X_t^2|\mathcal{G})\mathrm{E}(Z_t^2|\mathcal{G})]^{1/2}$ almost surely. Therefore

$$\begin{aligned}
& \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(Z_t^2 I(|Z_t| > \varepsilon\sqrt{n})|\mathcal{F}_{t-1}) \quad \text{(use the conditional Cauchy-Schwartz to split these terms)} \\
\leq\ & \frac{1}{n}\sum_{t=1}^{n} \left\{\mathrm{E}(X_{t-1}^4\varepsilon_t^4|X_{t-1})\mathrm{E}(I(|X_{t-1}\varepsilon_t| > \eta\sqrt{n})^2|X_{t-1})\right\}^{1/2} \\
\leq\ & \frac{1}{n}\sum_{t=1}^{n} X_{t-1}^2\mathrm{E}(\varepsilon_t^4)^{1/2}\left\{\mathrm{E}(I(|X_{t-1}\varepsilon_t| > \eta\sqrt{n})^2|X_{t-1})\right\}^{1/2}, \qquad (9.17)
\end{aligned}$$

almost surely. We note that rather than use the conditional Cauchy-Schwartz inequality we can use a generalisation of it called the conditional Hölder inequality. The Hölder inequality states that if $p^{-1} + q^{-1} = 1$, then $\mathrm{E}(XY|\mathcal{F}) \leq \{\mathrm{E}(X^p|\mathcal{F})\}^{1/p}\{\mathrm{E}(Y^q|\mathcal{F})\}^{1/q}$ almost surely. The advantage of using this inequality is that one can reduce the moment assumptions on X_t.

Returning to (9.17), and studying $\mathrm{E}(I(|X_{t-1}\varepsilon_t| > \varepsilon)^2|X_{t-1})$ we use that $\mathrm{E}(I(A)^2) = \mathrm{E}(I(A)) = \mathbb{P}(A)$ and the Chebyshev inequality to show

$$\begin{aligned}
\mathrm{E}\left(I(|X_{t-1}\varepsilon_t| > \eta\sqrt{n})^2|X_{t-1}\right) &= \mathrm{E}\left(I(|X_{t-1}\varepsilon_t| > \eta\sqrt{n})|X_{t-1}\right) \\
&= \mathrm{E}\left(I\left(|\varepsilon_t| > \frac{\eta\sqrt{n}}{X_{t-1}}\right)|X_{t-1}\right) \\
&= P_\varepsilon\left(|\varepsilon_t| > \frac{\eta\sqrt{n}}{X_{t-1}}\right) \leq \frac{X_{t-1}^2\mathrm{var}(\varepsilon_t)}{\eta^2 n}. \qquad (9.18)
\end{aligned}$$

Substituting (9.18) into (9.17) we have

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left(Z_t^2 I(|Z_t|>\eta\sqrt{n})|\mathcal{F}_{t-1}\right) &\leq \frac{1}{n}\sum_{t=1}^{n}X_{t-1}^2\mathrm{E}(\varepsilon_t^4)^{1/2}\left\{\frac{X_{t-1}^2\mathrm{var}(\varepsilon_t)}{\eta^2 n}\right\}^{1/2}\\ &\leq \frac{\mathrm{E}(\varepsilon_t^4)^{1/2}}{\eta n^{3/2}}\sum_{t=1}^{n}|X_{t-1}|^3\mathrm{E}(\varepsilon_t^2)^{1/2}\\ &\leq \frac{\mathrm{E}(\varepsilon_t^4)^{1/2}\mathrm{E}(\varepsilon_t^2)^{1/2}}{\eta n^{1/2}}\frac{1}{n}\sum_{t=1}^{n}|X_{t-1}|^3.\end{aligned}$$

If $\mathrm{E}(\varepsilon_t^4)<\infty$, then $\mathrm{E}(X_t^4)<\infty$, therefore by using the ergodic theorem we have $\frac{1}{n}\sum_{t=1}^{n}|X_{t-1}|^3 \stackrel{\text{a.s.}}{\to} \mathrm{E}(|X_0|^3)$. Since almost sure convergence implies convergence in probability we have

$$\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}(Z_t^2 I(|Z_t|>\eta\sqrt{n})|\mathcal{F}_{t-1}) \leq \underbrace{\frac{\mathrm{E}(\varepsilon_t^4)^{1/2}\mathrm{E}(\varepsilon_t^2)^{1/2}}{\eta n^{1/2}}}_{\to 0}\underbrace{\frac{1}{n}\sum_{t=1}^{n}|X_{t-1}|^3}_{\stackrel{\mathcal{P}}{\to}\mathrm{E}(|X_0|^3)} \stackrel{\mathcal{P}}{\to} 0.$$

Hence condition (9.13) is satisfied.

C3 : Finally, we need to verify that

$$\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}(Z_t^2|\mathcal{F}_{t-1}) \stackrel{\mathcal{P}}{\to} \sigma^2.$$

Since $\{X_t\}_t$ is an ergodic sequence we have

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}(Z_t^2|\mathcal{F}_{t-1}) = \frac{1}{n}\sum_{t=1}^{n}\mathrm{E}(X_{t-1}^2\varepsilon_t^2|X_{t-1})\\ = \;&\frac{1}{n}\sum_{t=1}^{n}X_{t-1}^2\mathrm{E}(\varepsilon_t^2|X_{t-1}) = \mathrm{E}(\varepsilon_t^2)\underbrace{\frac{1}{n}\sum_{t=1}^{n}X_{t-1}^2}_{\stackrel{\text{a.s.}}{\to}\mathrm{E}(X_0^2)} \stackrel{\text{a.s.}}{\to} \mathrm{E}(\varepsilon^2)\mathrm{E}(X_0^2) = \sigma^2\Sigma_1,\end{aligned}$$

hence we have verified condition (9.14).

Altogether conditions C1-C3 imply that

$$\sqrt{n}\nabla\mathcal{L}_n(\phi) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}X_{t-1}\varepsilon_t \stackrel{\mathcal{D}}{\to} \mathcal{N}(0,\sigma^2\Sigma_1). \tag{9.19}$$

Therefore

$$\sqrt{n}(\widehat{\phi}_n - \phi) = \underbrace{\left(\frac{1}{2}\nabla^2 \mathcal{L}_n\right)^{-1}}_{\stackrel{\text{a.s.}}{\to}(\mathrm{E}(X_0^2))^{-1}} \underbrace{\sqrt{n}S_n}_{\stackrel{\mathcal{D}}{\to}\mathcal{N}(0,\sigma^2 c(0))} \quad . \tag{9.20}$$

Using that $\mathrm{E}(X_0^2) = c(0)$, this implies that

$$\sqrt{n}(\hat{\phi}_n - \phi) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, \sigma^2 \Sigma_1^{-1}). \tag{9.21}$$

Thus we have derived the limiting distribution of $\hat{\phi}_n$.

Remark 9.1.7 *We recall that*

$$(\hat{\phi}_n - \phi) = -\left(\nabla^2 \mathcal{L}_n\right)^{-1} \nabla \mathcal{L}_n(\phi) = \frac{\frac{1}{n-1}\sum_{t=2}^n \varepsilon_t X_{t-1}}{\frac{1}{n-1}\sum_{t=2}^n X_{t-1}^2}, \tag{9.22}$$

and that $\mathrm{var}(\frac{1}{n-1}\sum_{t=2}^n \varepsilon_t X_{t-1}) = \frac{1}{n-1}\sum_{t=2}^n \mathrm{var}(\varepsilon_t X_{t-1}) = O(\frac{1}{n})$. *This implies*

$$(\hat{\phi}_n - \phi) = O_p(n^{-1/2}).$$

Indeed the results also holds almost surely

$$(\hat{\phi}_n - \phi) = O(n^{-1/2}). \tag{9.23}$$

The same result is true for autoregressive processes of arbitrary finite order. That is

$$\sqrt{n}(\hat{\underline{\phi}}_n - \underline{\phi}) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, \sigma^2 \Sigma_p^{-1}). \tag{9.24}$$

9.2 Estimation for ARMA models

Let us suppose that $\{X_t\}$ satisfies the ARMA representation

$$X_t - \sum_{i=1}^p \phi_i X_{t-i} = \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j},$$

and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ and $\sigma^2 = \text{var}(\varepsilon_t)$. We will suppose for now that p and q are known. The objective in this section is to consider various methods for estimating these parameters.

9.2.1 The Gaussian maximum likelihood estimator

We now derive the Gaussian maximum likelihood estimator (GMLE) to estimate the parameters $\underline{\theta}$ and $\underline{\phi}$. Let $\underline{X}_n' = (X_1, \ldots, X_n)$. The criterion (the GMLE) is constructed as if $\{X_t\}$ were Gaussian, but this need not be the case. The likelihood is similar to the likelihood given in (9.5), but just as in the autoregressive case it can be not directly maximised, i.e.

$$L_n(\phi, \theta, \sigma) \quad = \quad -\log\det(\Sigma_n(\phi, \theta, \sigma)) - \underline{X}_n'\Sigma_n(\phi, \theta, \sigma)^{-1}\underline{X}_n, \tag{9.25}$$

where $\Sigma_n(\phi, \theta, \sigma)$ the variance covariance matrix of $\underline{X}_n$. However, the above can be written in a tractable way by using conditioning

$$L_n(\zeta) \quad = \quad \log f(X_1; \zeta) + \sum_{t=1}^{n-1} \log f(X_{t+1}|X_t, \ldots, X_1; \zeta) \quad \text{(by repeated conditioning)}$$

where for simplicity we set $\zeta = (\theta, \phi, \sigma)$. Note that $f(X_{t+1}|X_1, \ldots, X_t, \zeta)$ is the conditional density of X_{t+1} given $X_1, \ldots, X_t$. Under the assumption the process is Gaussian, the conditional distribution of X_{t+1} conditioned on the past is Gaussian where

$$X_{t+1}|X_t, \ldots, X_1 \sim N\left(P_{X_1,\ldots,X_t}(X_{t+1}, \zeta), \text{var}[X_{t+1}|X_t, \ldots, X_1, \zeta]\right).$$

and $\text{var}[X_{t+1}|X_t, \ldots, X_1, \zeta] = \text{E}_\zeta[X_{t+1} - P_{X_1,\ldots,X_t}(X_{t+1}; \zeta)]^2$. We recall from Sections 7.5.1 and (7.8) and equation (7.5.1) that the coefficients of $P_{X_1,\ldots,X_t}(X_{t+1})$ are the best fitting AR(t) parameters based on the autocovariance $\{c(r; \zeta)\}_r$ which implictly depends on the ARMA parameters ζ. Thus

$$P_{X_1,\ldots,X_t}(X_{t+1}; \zeta) = X_{t+1|t} = \sum_{j=1}^{t} a_{j,t}(\zeta) X_{t+1-j}$$

and

$$\text{E}_\zeta[X_{t+1} - P_{X_1,\ldots,X_t}(X_{t+1}; \zeta)]^2 = r(t+1; \zeta) = c(0; \zeta) - \underline{r}_t(\zeta)'\Sigma_t(\zeta)^{-1}\underline{r}_t(\zeta).$$

The above looks cumbersome but it can be evaluated using the Levinson-Durbin algorithm. Further, if $t > \max(p, q)$, then by using equation (7.8)

$$X_{t+1|t}(\zeta) = \sum_{j=1}^{t} a_{j,t}(\zeta) X_{t+1-j} = \sum_{j=1}^{p} \phi_j(\zeta) X_{t+1-j} + \sum_{i=1}^{q} \theta_{t,i}(\zeta)(X_{t+1-i} - X_{t+1-i|t-i}(\zeta)).$$

Thus under Gaussianity

$$f(X_{t+1}|X_t, \ldots, X_1; \zeta) = \frac{1}{\sqrt{2\pi r(t+1;\zeta)}} \exp\left(-\frac{(X_{t+1} - X_{t+1|t}(\zeta))^2}{2r(t+1;\zeta)}\right)$$

and

$$\log f(X_{t+1}|X_t, \ldots, X_1; \zeta) \propto -\log r(t+1;\zeta) - \frac{(X_{t+1} - X_{t+1|t}(\zeta))^2}{r(t+1;\zeta)}.$$

Substituting this into $L_n(\zeta)$ gives

$$L_n(\zeta) = -\sum_{t=1}^{n} \log r(t;\zeta) - \frac{X_1^2}{r(1;\zeta)} - \sum_{t=1}^{n-1} \frac{(X_{t+1} - \sum_{j=1}^{t} \phi_{t+1,j}(\zeta) X_{t+1-j})^2}{r(t+1;\zeta)}.$$

An alternative but equivalent derivation of the above is to use the Cholesky decomposition of $\Sigma_n(\zeta)$ (see Section 7.5.3, equation (7.14)). For each set of parameters ζ and $r(t+1;\zeta)$ and $X_{t+1-i|t-i}(\zeta)$ can be evaluated. Thus the maximum likelihood estimators are the parameters $\widehat{\zeta}_n = \arg\max_\zeta L_n(\zeta)$.

The above can be a little difficult to evaluate. We give describe some popular approximations below.

9.2.2 The approximate Gaussian likelihood

We obtain an approximation to the log-likelihood which simplifies the estimation scheme. We recall in Section 7.8 we approximated $X_{t+1|t}$ with $\widehat{X}_{t+1|t}$. This motivates the approximation where we replace $X_{t+1|t}$ in $\widehat{L}_n(\zeta)$ with $\hat{X}_{t+1|t}$, where $\hat{X}_{t+1|t}$ is defined in (7.21)

$$\widehat{X}_{t+1|t} = \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i (X_{t+1-i} - \widehat{X}_{t-i}(1)) = \sum_{s=1}^{t} a_j X_{t+1-s} + \sum_{s=1}^{\max(p,q)} b_s X_s.$$

and $r(t;\zeta)$ with σ^2. This gives the approximate Gaussian log-likelihood

$$\begin{aligned}\widehat{L}_n(\zeta) &= -\sum_{t=1}^{n}\log\sigma^2 - \sum_{t=2}^{n}\frac{[X_t-\widehat{X}_{t|t-1}(\zeta)]^2}{\sigma^2} \\ &= -\sum_{t=1}^{n}\log\sigma^2 - \sum_{t=2}^{n}\frac{[(\theta(B)^{-1}\phi(B))_{[t]}X_t]^2}{\sigma^2}\end{aligned}$$

where $(\theta(B)^{-1}\phi(B))_{[t]}$ denotes the approximation of the polynomial in B to the tth order. The approximate likelihood greatly simplifies the estimation scheme because the derivatives (which is the main tool used in the maximising it) can be easily obtained. To do this we note that

$$\begin{aligned}\frac{d}{d\theta_i}\frac{\phi(B)}{\theta(B)}X_t &= -\frac{B^i\phi(B)}{\theta(B)^2}X_t = -\frac{\phi(B)}{\theta(B)^2}X_{t-i} \\ \frac{d}{d\phi_j}\frac{\phi(B)}{\theta(B)}X_t &= -\frac{B^j}{\theta(B)^2}X_t = -\frac{1}{\theta(B)^2}X_{t-j}\end{aligned} \tag{9.26}$$

therefore

$$\frac{d}{d\theta_i}\left(\frac{\phi(B)}{\theta(B)}X_t\right)^2 = -2\left(\frac{\phi(B)}{\theta(B)}X_t\right)\left(\frac{\phi(B)}{\theta(B)^2}X_{t-i}\right) \text{ and } \frac{d}{d\phi_j}\left(\frac{\phi(B)}{\theta(B)}X_t\right)^2 = -2\left(\frac{\phi(B)}{\theta(B)}X_t\right)\left(\frac{1}{\theta(B)^2}X_{t-j}\right). \tag{9.27}$$

Substituting this into the approximate likelihood gives the derivatives

$$\begin{aligned}\frac{\partial\widehat{L}}{\partial\theta_i} &= -\frac{2}{\sigma^2}\sum_{t=2}^{n}\left[\left(\theta(B)^{-1}\phi(B)\right)_{[t]}X_t\right]\left[\left(\frac{\phi(B)}{\theta(B)^2}\right)_{[t-i]}X_{t-i}\right] \\ \frac{\partial\widehat{L}}{\partial\phi_j} &= -\frac{2}{\sigma^2}\sum_{t=1}^{n}\left[\left(\theta(B)^{-1}\phi(B)\right)_{[t]}X_t\right]\left[\left(\frac{1}{\theta(B)}\right)_{[t-j]}X_{t-j}\right] \\ \frac{\partial\widehat{L}}{\partial\sigma^2} &= \frac{1}{\sigma^2} - \frac{1}{n\sigma^4}\sum_{t=1}^{n}\left[\left(\theta(B)^{-1}\phi(B)\right)_{[t]}X_t\right]^2.\end{aligned} \tag{9.28}$$

We then use the Newton-Raphson scheme to maximise the approximate likelihood.

It should be mentioned that such approximations are very common in time series, though as with all approximation, they can be slightly different. Lütkepohl (2005), Section 12.2 gives a very similar approximation, but uses the variance/covariance matrix of the time series $\underline{X}_n$ as the basis of the approximation. By approximating the variance/covariance he finds an approximation of the likelihood.

9.2.3 Estimation using the Kalman filter

We now state another approximation using the Kalman filter. We recall from Section 7.9 that the ARMA model satisfies the recursion

$$\begin{pmatrix} X(t+1|t+1) \\ X(t+2|t+1) \\ X(t+3|t) \\ \vdots \\ X(t+m-1|t+1) \\ X(t+m|t+1) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ \phi_m & \phi_{m-1} & \phi_{m-2} & \dots & \phi_2 & \phi_1 \end{pmatrix} \begin{pmatrix} X(t|t) \\ X(t+1|t) \\ X(t+2|t) \\ \vdots \\ X(t+m-2|t) \\ X(t+m-1|t) \end{pmatrix} + \varepsilon_{t+1} \begin{pmatrix} 1 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{m-2} \\ \psi_{m-1} \end{pmatrix}$$

$$\underline{Z}_{t+1} = F\underline{Z}_t + \underline{V}_{t+1}$$

for $t \in \mathbb{Z}$. The observation model is the ARMA(p, q) process where

$$X_{t+1} = (1, 0, \dots, 0)\underline{Z}_{t+1} \text{ where } t \in \mathbb{Z}.$$

$Q(\zeta) = \text{var}(\varepsilon_t)\underline{\psi}_m'\underline{\psi}_m$, $R = 0$, $H = (1, 0, \dots, 0)$

The Kalman equations

(1) Start with an initial value $\underline{Z}_{0|0}$. This part is where the approximation comes into play because Y_0 is not observed. This part causes me quite a lot of trauma because it means the likelihood based on the Kalman equations is also an approximation (though it rarely is stated that it is). Typically we set $\underline{Z}_{0|0} = (0, \dots, 0)$ and recommendations for $P_{0|0}$ are given in Jones (1980) and Akaiki (1978). Then for $t > 0$ iterate on steps (2) and (3) below.

(2) Prediction step

$$\widehat{Z}_{t+1|t}(\zeta) = F(\zeta)\widehat{Z}_{t|t}(\zeta)$$

and the corresponding mean squared error

$$P_{t+1|t}(\zeta) = F(\zeta)P_{t|t}F(\zeta)^* + Q(\zeta).$$

(3) Update step The conditional expectation

$$\widehat{Z}_{t+1|t+1}(\zeta) = \widehat{Z}_{t+1|t}(\zeta) + K_{t+1}(\zeta)\left(X_{t+1} - H\widehat{Z}_{t+1|t}(\zeta)\right).$$

where

$$K_{t+1}(\zeta) = \frac{P_{t+1|t}(\zeta)H^*}{HP_{t+1|t}(\zeta)H^*}$$

and the corresponding mean squared error

$$P_{t+1|t+1}(\zeta) = P_{t+1|t}(\zeta) - K_t(\zeta)HP_{t+1|t}(\zeta) = (I - K_t(\zeta)H)P_{t+1|t}(\zeta).$$

Thus $\widehat{Z}_{t+1|t}(\zeta)_{(1)} \approx X_{t+1|t}(\zeta)$ and $P_{t+1|t}(\zeta) \approx r(t+1|\zeta)$ thus an approximation of the Gaussian likeilhood is

$$\widehat{L}_n(\zeta) = -\sum_{t=1}^{n} \log P_{t|t-1}(\zeta) - \sum_{t=1}^{n} \frac{(X_t - \widehat{Z}_{t|t-1}(\zeta)_{(1)})^2}{P_{t|t-1}(\zeta)}.$$

In order to maximise the above (using the Newton Raphson scheme) the derivative of the above needs to be evaluated with respect to ζ. How exactly this is done is not so clear to me. $\widehat{Z}_{t|t-1}(\zeta)$ could be evaluated by defining a new set of state space equations for $(\underline{Z}_{t+1}(\zeta), \nabla_\zeta \underline{Z}_{t+1}(\zeta))$ but I not so sure how this can be done for $P_{t|t-1}(\zeta)$. Nevertheless somehow it is done successfully used to maximise the likelihood.

9.2.4 Sampling properties of the ARMA maximum likelihood estimator

It can be shown that the approximate likelihood is close to the actual true likelihood and asymptotically both methods are equivalent.

Theorem 9.2.1 *Let us suppose that X_t has a causal and invertible ARMA representation*

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and $\mathrm{var}[\varepsilon_t] = \sigma^2$ (we do not assume Gaussian-

ity). Then the (quasi)-Gaussian

$$\sqrt{n}\left(\begin{array}{c}\hat{\underline{\phi}}_n-\underline{\phi}\\ \hat{\underline{\theta}}_n-\underline{\theta}\end{array}\right)\stackrel{\mathcal{D}}{\to}\mathcal{N}(0,\sigma^2\Lambda^{-1}),$$

with

$$\Lambda=\left(\begin{array}{cc}\mathrm{E}(\mathbf{U}_t\mathbf{U}_t') & \mathrm{E}(\mathbf{V}_t\mathbf{U}_t')\\ \mathrm{E}(\mathbf{U}_t\mathbf{V}_t') & \mathrm{E}(\mathbf{V}_t\mathbf{V}_t')\end{array}\right)$$

and $\mathbf{U}_t=(U_t,\ldots,U_{t-p+1})$ *and* $\mathbf{V}_t=(V_t,\ldots,V_{t-q+1})$*, where* $\{U_t\}$ *and* $\{V_t\}$ *are autoregressive processes which satisfy* $\phi(B)U_t=\varepsilon_t$ *and* $\theta(B)V_t=\varepsilon_t$*.*

We do not give the proof in this section, however it is possible to understand where this result comes from. We recall that that the maximum likelihood and the approximate likelihood are asymptotically equivalent. They are both approximations of the unobserved likelihood

$$\widetilde{L}_n(\boldsymbol{\theta}) \;=\; -\sum_{t=1}^{n}\log\sigma^2-\sum_{t=2}^{n-1}\frac{[X_{t+1}-X_t(1;\boldsymbol{\theta})]^2}{\sigma^2}=-\sum_{t=1}^{n}\log\sigma^2-\sum_{t=2}^{n-1}\frac{[\theta(B)^{-1}\phi(B)X_{t+1}]^2}{\sigma^2},$$

where $\boldsymbol{\theta}=(\phi,\theta,\sigma^2)$. This likelihood is infeasible in the sense that it cannot be maximised since the finite past $X_0,X_1,\ldots$ is unobserved, however is a very convenient tool for doing the asymptotic analysis. Using Lemma 7.8.1 we can show that all three likelihoods L_n, $\widehat{L}_n$ and $\widetilde{L}_n$ are all asymptotically equivalent. Therefore, to obtain the asymptotic sampling properties of L_n or $\widehat{L}_n$ we can simply consider the unobserved likelihood $\widetilde{L}_n$.

To show asymptotic normality (we assume here that the estimators are consistent) we need to consider the first and second derivative of $\widetilde{L}_n$ (since the asymptotic properties are determined by Taylor expansions). In particular we need to consider the distribution of $\frac{\partial\widetilde{L}_n}{\partial\boldsymbol{\theta}}$ at its true parameters and the expectation of $\frac{\partial^2\widetilde{L}_n}{\partial\boldsymbol{\theta}^2}$ at it's true parameters. We note that by using (9.27) we have

$$\begin{aligned}\frac{\partial\widetilde{L}}{\partial\theta_i} &= -\frac{2}{\sigma^2}\sum_{t=1}^{n}\left[\left(\theta(B)^{-1}\phi(B)\right)X_t\right]\left[\left(\frac{\phi(B)}{\theta(B)^2}\right)X_{t-i}\right]\\ \frac{\partial\widetilde{L}}{\partial\phi_j} &= -\frac{2}{\sigma^2}\sum_{t=1}^{n}\left[\left(\theta(B)^{-1}\phi(B)\right)X_t\right]\left[\left(\frac{1}{\theta(B)}\right)X_{t-j}\right]\end{aligned}\tag{9.29}$$

Since we are considering the derivatives at the true parameters we observe that $\left(\theta(B)^{-1}\phi(B)\right)X_t=$

ε_t,

$$\frac{\phi(B)}{\theta(B)^2}X_{t-i} = \frac{\phi(B)}{\theta(B)^2}\frac{\theta(B)}{\phi(B)}\varepsilon_{t-i} = \frac{1}{\theta(B)}\varepsilon_{t-i} = V_{t-i}$$

and

$$\frac{1}{\theta(B)}X_{t-j} = \frac{1}{\theta(B)}\frac{\theta(B)}{\phi(B)}\varepsilon_{t-j} = \frac{1}{\phi(B)}\varepsilon_{t-j} = U_{t-j}.$$

Thus $\phi(B)U_t = \varepsilon_t$ and $\theta(B)V_t = \varepsilon_t$ are autoregressive processes (compare with theorem). This means that the derivative of the unobserved likelihood can be written as

$$\frac{\partial \widetilde{L}}{\partial \theta_i} = -\frac{2}{\sigma^2}\sum_{t=1}^{n}\varepsilon_t U_{t-i} \text{ and } \frac{\partial \widetilde{L}}{\partial \phi_j} = -\frac{2}{\sigma^2}\sum_{t=1}^{n}\varepsilon_t V_{t-j} \tag{9.30}$$

Note that by causality ε_t, U_{t-i} and V_{t-j} are independent. Again like many of the other estimators we have encountered this sum is 'mean-like' so can show normality of it by using a central limit theorem designed for dependent data. Indeed we can show asymptotically normality of $\{\frac{\partial \widetilde{L}}{\partial \theta_i}; i = 1, \ldots, q\}$, $\{\frac{\partial \widetilde{L}}{\partial \phi_j}; j = 1, \ldots, p\}$ and their linear combinations using the Martingale central limit theorem, see Theorem 3.2 (and Corollary 3.1), Hall and Heyde (1980) - note that one can also use m-dependence. Moreover, it is relatively straightforward to show that $n^{-1/2}(\frac{\partial \widetilde{L}}{\partial \theta_i}, \frac{\partial \widetilde{L}}{\partial \phi_j})$ has the limit variance matrix Δ. Finally, by taking second derivative of the likelihood we can show that $\mathrm{E}[n^{-1}\frac{\partial^2 \widehat{L}}{\partial \boldsymbol{\theta}^2}] = \Delta$. Thus giving us the desired result.

9.2.5 The Hannan-Rissanen AR(∞) expansion method

The methods detailed above require good initial values in order to begin the maximisation (in order to prevent convergence to a local maximum).

We now describe a simple method first propose in Hannan and Rissanen (1982) and An et al. (1982). It is worth bearing in mind that currently the 'large p small n problem' is a hot topic. These are generally regression problems where the sample size n is quite small but the number of regressors p is quite large (usually model selection is of importance in this context). The methods proposed by Hannan involves expanding the ARMA process (assuming invertibility) as an $AR(\infty)$ process and estimating the parameters of the $AR(\infty)$ process. In some sense this can be considered as a regression problem with an infinite number of regressors. Hence there are some parallels

between the estimation described below and the 'large p, small n problem'.

As we mentioned in Lemma 4.10.1, if an ARMA process is invertible it is can be represented as

$$X_t = \sum_{j=1}^{\infty} b_j X_{t-j} + \varepsilon_t. \tag{9.31}$$

The idea behind Hannan's method is to estimate the parameters $\{b_j\}$, then estimate the innovations ε_t, and use the estimated innovations to construct a multiple linear regression estimator of the ARMA paramters $\{\theta_i\}$ and $\{\phi_j\}$. Of course in practice we cannot estimate all parameters $\{b_j\}$ as there are an infinite number of them. So instead we do a type of sieve estimation where we only estimate a finite number and let the number of parameters to be estimated grow as the sample size increases. We describe the estimation steps below:

(i) Suppose we observe $\{X_t\}_{t=1}^n$. Recalling (9.31), will estimate $\{b_j\}_{j=1}^{p_n}$ parameters. We will suppose that $p_n \to \infty$ as $n \to \infty$ and $p_n << n$ (we will state the rate below).

We use Yule-Walker to estimate $\{b_j\}_{j=1}^{p_n}$, where

$$\hat{\underline{b}}_{p_n} = \hat{\Sigma}_{p_n}^{-1} \hat{\underline{r}}_{p_n},$$

where

$$(\hat{\Sigma}_{p_n})_{i,j} = \frac{1}{n} \sum_{t=1}^{n-|i-j|} (X_t - \bar{X})(X_{t+|i-j|} - \bar{X}) \text{ and } (\hat{\underline{r}}_{p_n})_j = \frac{1}{n} \sum_{t=1}^{n-|j|} (X_t - \bar{X})(X_{t+|j|} - \bar{X}).$$

(ii) Having estimated the first $\{b_j\}_{j=1}^{p_n}$ coefficients we estimate the residuals with

$$\tilde{\varepsilon}_t = X_t - \sum_{j=1}^{p_n} \hat{b}_{j,n} X_{t-j}.$$

(iii) Now use as estimates of $\underline{\phi}_0$ and $\underline{\theta}_0$ $\tilde{\underline{\phi}}_n, \tilde{\underline{\theta}}_n$ where

$$\tilde{\underline{\phi}}_n, \tilde{\underline{\theta}}_n = \arg\min \sum_{t=p_n+1}^{n} (X_t - \sum_{j=1}^{p} \phi_j X_{t-j} - \sum_{i=1}^{q} \theta_i \tilde{\varepsilon}_{t-i})^2.$$

We note that the above can easily be minimised. In fact

$$(\widetilde{\underline{\phi}}_n, \widetilde{\underline{\theta}}_n) = \widetilde{\mathcal{R}}_n^{-1} \widetilde{\underline{s}}_n$$

where

$$\tilde{\mathcal{R}}_n = \frac{1}{n}\sum_{t=\max(p,q)}^{n} \widetilde{\underline{Y}}_t \widetilde{\underline{Y}}_t' \quad \text{and} \quad \widetilde{\underline{s}}_n = \frac{1}{n}\sum_{t=\max(p,q)}^{n} \widetilde{\underline{Y}}_t X_t,$$

$\widetilde{\underline{Y}}_t' = (X_{t-1}, \ldots, X_{t-p}, \widetilde{\varepsilon}_{t-1}, \ldots, \widetilde{\varepsilon}_{t-q})$.

9.3 The quasi-maximum likelihood for ARCH processes

In this section we consider an estimator of the parameters $\underline{a}_0 = \{a_j : j = 0, \ldots, p\}$ given the observations $\{X_t : t = 1, \ldots, N\}$, where $\{X_t\}$ is a ARCH(p) process. We use the conditional log-likelihood to construct the estimator. We will assume throughout that $\mathrm{E}(Z_t^2) = 1$ and $\sum_{j=1}^{p} \alpha_j = \rho < 1$.

We now construct an estimator of the ARCH parameters based on $Z_t \sim \mathcal{N}(0,1)$. It is worth mentioning that despite the criterion being constructed under this condition it is not necessary that the innovations Z_t are normally distributed. In fact in the case that the innovations are not normally distributed but have a finite fourth moment the estimator is still good. This is why it is called the quasi-maximum likelihood , rather than the maximum likelihood (similar to the how the GMLE estimates the parameters of an ARMA model regardless of whether the innovations are Gaussian or not).

Let us suppose that Z_t is Gaussian. Since $Z_t = X_t/\sqrt{a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2}$, $\mathrm{E}(X_t|X_{t-1}, \ldots, X_{t-p}) = 0$ and $\mathrm{var}(X_t|X_{t-1}, \ldots, X_{t-p}) = a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2$, then the log density of X_t given $X_{t-1}, \ldots, X_{t-p}$ is

$$\log(a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2) + \frac{X_t^2}{a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2}.$$

Therefore the conditional log density of $X_{p+1}, X_{p+2}, \ldots, X_n$ given $X_1, \ldots, X_p$ is

$$\sum_{t=p+1}^{n} \left(\log(a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2) + \frac{X_t^2}{a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2} \right).$$

This inspires the the conditional log-likelihood

$$\mathcal{L}_n(\underline{\alpha}) = \frac{1}{n-p}\sum_{t=p+1}^{n} \left(\log(\alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2) + \frac{X_t^2}{\alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2} \right).$$

To obtain the estimator we define the parameter space

$$\Theta = \{\underline{\alpha} = (\alpha_0, \ldots, \alpha_p) : \sum_{j=1}^{p} \alpha_j \leq 1, 0 < c_1 \leq \alpha_0 \leq c_2 < \infty, c_1 \leq \alpha_j\}$$

and assume the true parameters lie in its interior $\underline{a} = (a_0, \ldots, a_p) \in Int(\Theta)$. We let

$$\hat{\underline{a}}_n = \arg\min_{\underline{\alpha} \in \Theta} \mathcal{L}_n(\underline{\alpha}). \tag{9.32}$$

The method for estimation of GARCH parameters parallels the approximate likelihood ARMA estimator given in Section 9.2.1.

Exercise 9.2 *The objective of this question is to estimate the parameters of a random autoregressive process of order one*

$$X_t = (\phi + \xi_t) X_{t-1} + \varepsilon_t,$$

where, $|\phi| < 1$ *and* $\{\xi_t\}_t$ *and* $\{\varepsilon_t\}_t$ *are zero mean iid random variables which are independent of each other, with* $\sigma_\xi^2 = \text{var}[\xi_t]$ *and* $\sigma_\varepsilon^2 = \text{var}[\varepsilon_t]$.

Suppose that $\{X_t\}_{t=1}^n$ *is observed. We will assume for parts (a-d) that* ξ_t *and* ε_t *are Gaussian random variables. In parts (b-c) the objective is to construct an initial value estimator which is easy to obtain but not optimal in (d) to obtain the maximum likelihood estimator.*

(a) What is the conditional expectation (best predictor) of X_t *given the past?*

(b) Suppose that $\{X_t\}_{t=1}^n$ *is observed. Use your answer in part (a) to obtain an explicit expression for estimating* ϕ.

(c) Define residual as $\xi_t X_{t-1} + \varepsilon_t$. *Use your estimator in (b) to estimate the residuals.*

Evaluate the variance of $\xi_t X_{t-1} + \varepsilon_t$ *conditioned on* X_{t-1}. *By using the estimated residuals explain how the conditional variance can be used to obtain an explicit expression for estimating* σ_ξ^2 *and* σ_ε^2.

(d) By conditioning on X_1 *obtain the log-likelihood of* $X_2, \ldots, X_n$ *under the assumption of Guassianity of* ξ_t *and* ε_t. *Explain the role that (b) and (c) plays in your maximisation algorithm.*

(e) **Bonus question (only attempt if you really want to)**

Show that the expectation of the conditional log-likelihood is maximised at the true parameters *($\phi_0, \sigma^2_{0,\xi}$ and $\sigma^2_{0,\varepsilon}$) even when ξ_t and ε_t are not Gaussian.*

Hint: You may want to use that the function $g(x) = -\log x + x$ is minimum at $x = 1$ where $g(1) = 1$ and let

$$x = \frac{\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X^2_{t-1}}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X^2_{t-1}}.$$

Chapter 10

Spectral Representations

Prerequisites

- Knowledge of complex numbers.
- Have some idea of what the covariance of a complex random variable (we do define it below).
- Some idea of a Fourier transform (a review is given in Section A.3).
- The very useful result on the discrete Fourier transform:

$$\sum_{t=1}^{n} \exp\left(it\frac{2\pi j}{n}\right) = \begin{cases} 0 & j \neq n\mathbb{Z} \\ n & j \in \mathbb{Z} \end{cases} . \tag{10.1}$$

Objectives

- Know the definition of the spectral density.
- The spectral density is always non-negative and this is a way of checking that a sequence is actually non-negative definite (is a autocovariance).
- The DFT of a second order stationary time series is almost uncorrelated.
- The spectral density of an ARMA time series, and how the roots of the characteristic polynomial of an AR may influence the spectral density function.
- There is no need to understand the proofs of either Bochner's (generalised) theorem or the spectral representation theorem, just know what these theorems are. However, you should

know the proof of Bochner's theorem in the simple case that $\sum_r |rc(r)| < \infty$.

10.1 How we have used Fourier transforms so far

We recall in Section 2.5 that we considered models of the form

$$X_t = A\cos(\omega t) + B\sin(\omega t) + \varepsilon_t \qquad t = 1, \ldots, n. \tag{10.2}$$

where ε_t are iid random variables with mean zero and variance σ^2 and ω is unknown. We estimated the frequency ω by taking the Fourier transform $J_n(\omega) = \frac{1}{\sqrt{n}}\sum_{t=1}^n X_t e^{it\omega}$ and using as an estimator of ω, the value which maximised $|J_n(\omega)|^2$. As the sample size grows the peak (which corresponds the frequency estimator) grows in size. Besides the fact that this corresponds to the least squares estimator of ω, we note that

$$\begin{aligned} \frac{1}{\sqrt{n}}J_n(\omega_k) &= \frac{1}{2\pi n}\sum_{t=1}^n X_t \exp(it\omega_k) \\ &= \underbrace{\frac{1}{2\pi n}\sum_{t=1}^n \mu(\frac{t}{n})\exp(it\omega_k)}_{=O(1)} + \underbrace{\frac{1}{2\pi n}\sum_{t=1}^n \varepsilon_t \exp(it\omega_k)}_{=O_p(n^{-1/2}) \text{ compare with } \frac{1}{n}\sum_{t=1}^n \varepsilon_t} \end{aligned} \tag{10.3}$$

where $\omega_k = \frac{2\pi k}{n}$, is an estimator the the Fourier transform of the deterministic mean at frequency k. In the case that the mean is simply the sin function, there is only one frequency which is non-zero. A plot of one realization ($n = 128$), periodogram of the realization, periodogram of the iid noise and periodogram of the sin function is given in Figure 10.1. Take careful note of the scale (y-axis), observe that the periodogram of the sin function dominates the the periodogram of the noise (magnitudes larger). We can understand why from (10.3), where the asymptotic rates are given and we see that the periodogram of the deterministic signal is estimating $n\times$Fourier coefficient, whereas the periodgram of the noise is $O_p(1)$. However, this is an asymptotic result, for small samples sizes you may not see such a big difference between deterministic mean and the noise. Next look at the periodogram of the noise we see that it is very erratic (we will show later that this is because it is an **inconsistent** estimator of the spectral density function), however, despite the erraticness, the amount of variation overall frequencies seems to be same (there is just one large peak - which could be explained by the randomness of the periodogram).

Returning again to Section 2.5, we now consider the case that the sin function has been cor-

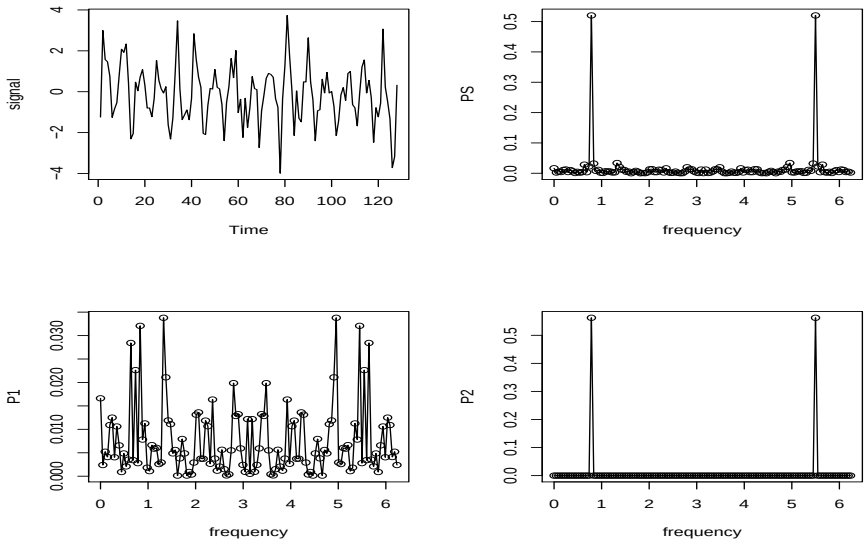


Figure 10.1: Top Left: Realisation of (2.18) ($2\sin(\frac{2\pi t}{8})$) with iid noise, Top Right: Periodogram of sin + noise. Bottom Left: Periodogram of just the noise. Bottom Right: Periodogram of just the sin function.

rupted by colored noise, which follows an AR(2) model

$$\varepsilon_t = 1.5\varepsilon_{t-1} - 0.75\varepsilon_{t-2} + \epsilon_t. \tag{10.4}$$

A realisation and the corresponding periodograms are given in Figure 10.2. The results are different to the iid case. The peak in the periodogram no longer corresponds to the period of the sin function. From the periodogram of the just the AR(2) process we observe that it erratic, just as in the iid case, however, there appears to be varying degrees of variation over the frequencies (though this is not so obvious in this plot). We recall from Chapters 2 and 3, that the AR(2) process has a pseudo-period, which means the periodogram of the colored noise will have pronounced peaks which correspond to the frequencies around the pseudo-period. It is these pseudo-periods which are dominating the periodogram, which is giving a peak at frequency that does not correspond to the sin function. However, asymptotically the rates given in (10.3) still hold in this case too. In other words, for large enough sample sizes the DFT of the signal should dominate the noise. To see that this is the case, we increase the sample size to $n = 1024$, a realisation is given in Figure 10.3. We see that the period corresponding the sin function dominates the periodogram. Studying the periodogram of just the AR(2) noise we see that it is still erratic (despite the large sample size),

but we also observe that the variability clearly changes over frequency.

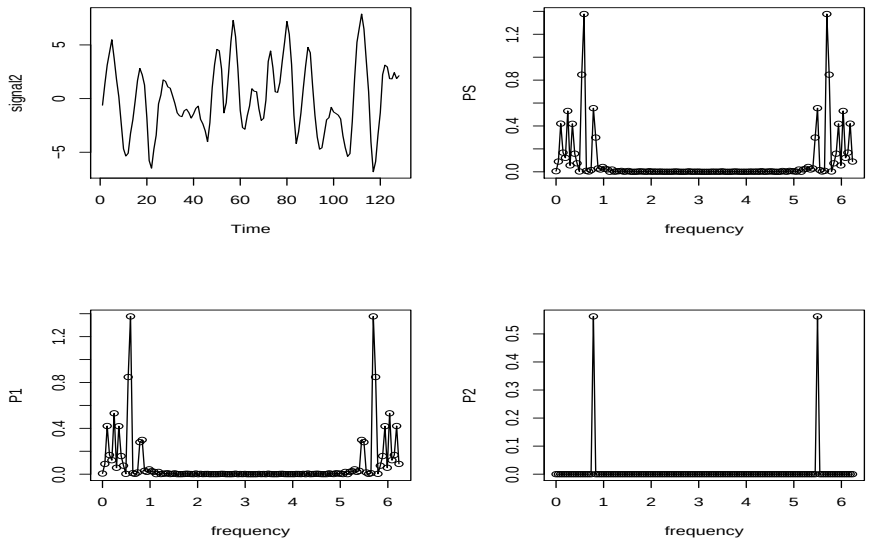


Figure 10.2: Top Left: Realisation of (2.18) ($2\sin(\frac{2\pi t}{8})$) with AR(2) noise ($n = 128$), Top Right: Periodogram. Bottom Left: Periodogram of just the AR(2) noise. Bottom Right: Periodogram of the sin function.

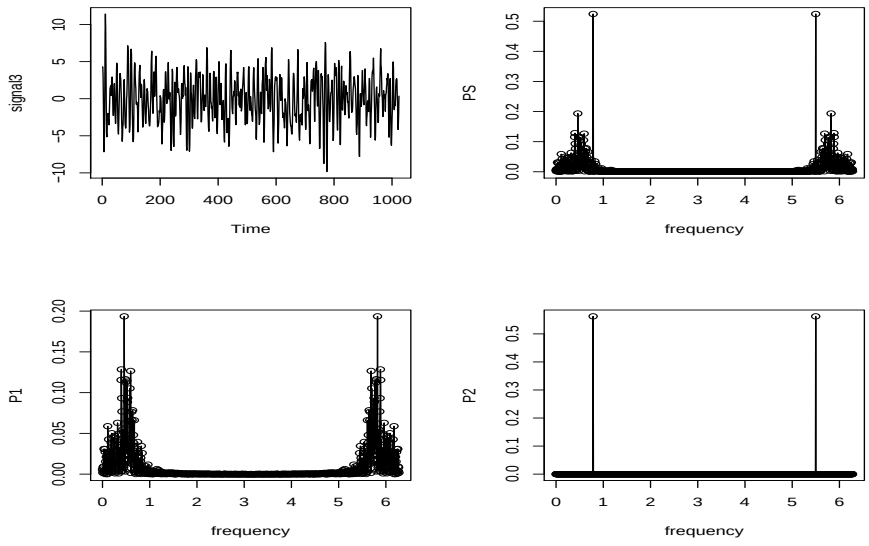


Figure 10.3: Top Left: Realisation of (2.18) ($2\sin(\frac{2\pi t}{8})$) with AR(2) noise ($n = 1024$), Top Right: Periodogram. Bottom Left: Periodogram of just the AR(2) noise. Bottom Right: Periodogram of the sin function.

From now on we focus on the constant mean stationary time series (eg. iid noise and the AR(2))

(where the mean is either constant or zero). As we have observed above, the periodogram is the absolute square of the discrete Fourier Transform (DFT), where

$$J_n(\omega_k) \quad = \quad \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_t \exp(it\omega_k). \tag{10.5}$$

This is simply a (linear) transformation of the data, thus it easily reversible by taking the inverse DFT

$$X_t = \frac{\sqrt{2\pi}}{\sqrt{n}} \sum_{t=1}^{n} J_n(\omega_k) \exp(-it\omega_k). \tag{10.6}$$

Therefore, just as one often analyzes the log transform of data (which is also an invertible transform), one can analyze a time series through its DFT.

In Figure 10.4 we give plots of the periodogram of an iid sequence and AR(2) process defined in equation (10.4). We recall from Chapter 3, that the periodogram is an **inconsistent** estimator of the spectral density function $f(\omega) = (2\pi)^{-1} \sum_{r=-\infty}^{\infty} c(r) \exp(ir\omega)$ and a plot of the spectral density function corresponding to the iid and AR(2) process defined in (**??**). We will show later that by inconsistent estimator we mean that $\mathrm{E}[|J_n(\omega_k)|^2] = f(\omega_k) + O(n^{-1})$ but $\mathrm{var}[|J_n(\omega_k)|^2] \not\to 0$ as $n \to \infty$. this explains why the general 'shape' of $|J_n(\omega_k)|^2$ looks like $f(\omega_k)$ but $|J_n(\omega_k)|^2$ is extremely erratic and variable.

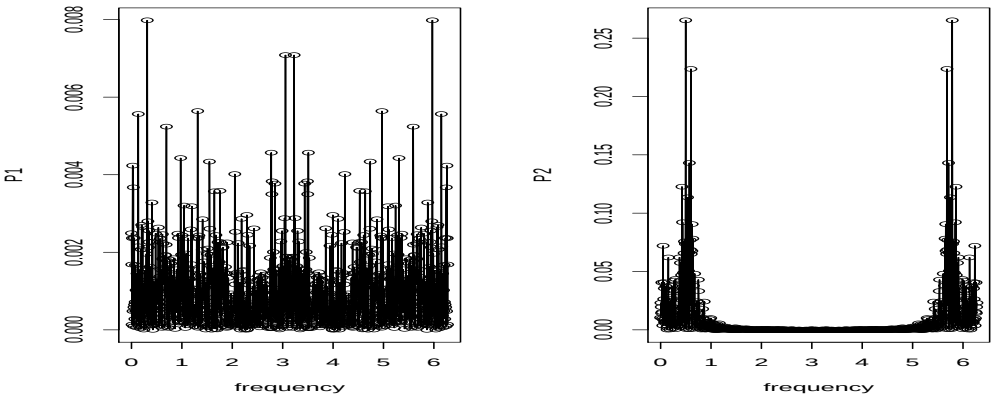


Figure 10.4: Left: Periodogram of iid noise. Right: Periodogram of AR(2) process.

Remark 10.1.1 (Properties of the spectral density function) *The spectral density function was first introduced in in Section 3.4. We recall that given an autoregressive process* $\{c(k)\}$*, the*

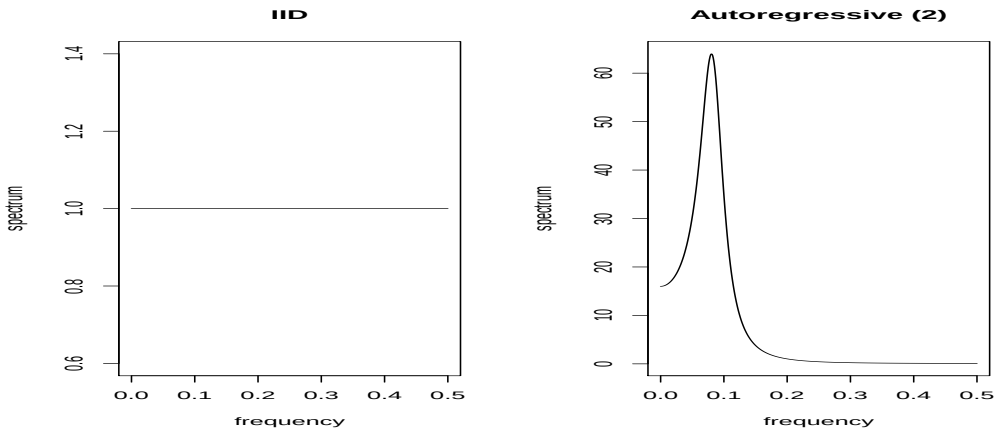


Figure 10.5: Left: Spectral density of iid noise. Right: Spectral density of AR(2), note that the interval $[0, 1]$ corresponds to $[0, 2\pi]$ in Figure 10.5

spectral density is defined as

$$f(\omega) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} c(r) \exp(2\pi i r).$$

And visa versa, given the spectral density we can recover the autocovariance via the inverse transform $c(r) = \int_0^{2\pi} f(\omega) \exp(-2\pi i r\omega) d\omega$. *We recall from Section 3.4 that the spectral density function can be used to construct a valid autocovariance function since only a sequence whose Fourier transform is real and positive can be positive definite.*

In Section 7.8 we used the spectral density function to define conditions under which the variance covariance matrix of a stationary time series had minimum and maximim eigenvalues. Now from the discussion above we observe that the variance of the DFT is approximately the spectral density function (note that for this reason the spectral density is sometimes called the power spectrum).

We now collect some of the above observations, to summarize some of the basic properties of the DFT:

(i) We note that $\overline{J_n(\omega_k)} = J_n(\omega_{n-k})$, therefore, all the information on the time series is contain in the first $n/2$ frequencies $\{J_n(\omega_k); k = 1, \ldots, n/2\}$.

(ii) If the time series $\mathrm{E}[X_t] = \mu$ and $k \neq 0$ then

$$\mathrm{E}[J_n(\omega_k)] = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mu \exp(it\omega_k) = 0.$$

If $k = 0$ then

$$\mathrm{E}[J_n(\omega_0)] = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mu = \sqrt{n}\mu.$$

In other words, the mean of the DFT (at non-zero frequencies) is zero regardless of whether the time series has a zero mean (it just needs to have a constant mean).

(iii) However, unlike the original stationary time series, we observe that the variance of the DFT depends on frequency (unless it is a white noise process) and that for $k \neq 0$, $\mathrm{var}[J_n(\omega_k)] = \mathrm{E}[|J_n(\omega_k)|^2] = f(\omega_k) + O(n^{-1})$.

The focus of this chapter will be on properties of the spectral density function (proving some of the results we stated previously) and on the so called Cramer representation (or spectral representation) of a second order stationary time series. However, before we go into these results (and proofs) we give one final reason why the analysis of a time series is frequently done by transforming to the frequency domain via the DFT. Above we showed that there is a one-to-one correspondence between the DFT and the original time series, below we show that the DFT almost decorrelates the stationary time series. In other words, one of the main advantages of working within the frequency domain is that we have transformed a correlated time series into something that it almost uncorrelated (this also happens to be a heuristic reason behind the spectral representation theorem).

10.2 The 'near' uncorrelatedness of the DFT

Let $\underline{X}_n = \{X_t; t = 1, \ldots, n\}$ and $\Sigma_n = \mathrm{var}[\underline{X}_n]$. It is clear that $\Sigma_n^{-1/2}\underline{X}_n$ is an uncorrelated sequence. This means to formally decorrelate $\underline{X}_n$ we need to know $\Sigma_n^{-1/2}$. However, if X_t is a second order stationary time series, something curiously, remarkable happens. The DFT, almost uncorrelates the $\underline{X}_n$. The implication of this is extremely useful in time series, and we shall be using this transform in estimation in Chapter 11.

We start by defining the Fourier transform of $\{X_t\}_{t=1}^n$ as

$$J_n(\omega_k) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_t \exp(ik\frac{2\pi t}{n}) \qquad k = 1, \ldots, n,$$

where the frequences $\omega_k = 2\pi k/n$ are often called the fundamental, Fourier frequencies. Note that

in R $\sum_{t=1}^{n} X_t \exp(-ik\frac{2\pi t}{n})$ is evaluated with the `fft` function for $k = 0, 1 \ldots, n-1$, where we observe that $k = 0$ is the same as $k = n$. To evaluate $\sum_{t=1}^{n} X_t \exp(ik\frac{2\pi t}{n})$ for $k = 0, 1, \ldots, n-1$ one needs to use the function `fft(x,inverse = TRUE)`. Keep in mind that

$$J_n(\omega_k) = \overline{J_n(\omega_{n-k})},$$

so it does not matter which definition one uses.

Below we state some of its properties.

Lemma 10.2.1 *Suppose $\{X_t\}$ is a second order stationary time series, where $\sum_r |rc(r)| < \infty$. Then we have*

$$\operatorname{cov}(J_n(\frac{2\pi k_1}{n}), J_n(\frac{2\pi k_2}{n})) = \begin{cases} f(\frac{2\pi k}{n}) + O(\frac{1}{n}) & k_1 = k_2 \\ O(\frac{1}{n}) & 1 \le k_1 \ne k_2 \le n/2 \end{cases}$$

where $f(\omega) = \frac{1}{2\pi}\sum_{r=-\infty}^{\infty} c(r)\exp(ir\omega)$. If one wants to consider the real and imaginary parts of $J_n(\omega_k)$ then

$$\operatorname{cov}(J_{n,C}(\frac{2\pi k_1}{n}), J_{n,C}(\frac{2\pi k_2}{n})) = \begin{cases} f(\frac{2\pi k}{n}) + O(\frac{1}{n}) & k_1 = k_2 \\ O(\frac{1}{n}) & 1 \le k_1 \ne k_2 \le n/2 \end{cases}$$

$$\operatorname{cov}(J_{n,S}(\frac{2\pi k_1}{n}), J_{n,S}(\frac{2\pi k_2}{n})) = \begin{cases} f(\frac{2\pi k}{n}) + O(\frac{1}{n}) & k_1 = k_2 \\ O(\frac{1}{n}) & 1 \le k_1 \ne k_2 \le n/2 \end{cases}$$

and $\operatorname{cov}[J_{n,C}(\frac{2\pi k_1}{n}), J_{n,S}(\frac{2\pi k_2}{n})] = O(n^{-1})$ for $1 \le k_1, k_2 \le n/2$, where

$$J_{n,C}(\omega_k) = \frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n} X_t \cos(t\omega_k), \quad J_{n,S}(\omega_k) = \frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n} X_t \sin(t\omega_k).$$

We prove the result in Section 10.2.2.

10.2.1 Testing for second order stationarity: An application of the near decorrelation property

We evaluate the DFT using the following piece of code (note that we do not standardize by $\sqrt{2\pi}$)

```
dft <- function(x){
            n=length(x)
            dft <- fft(x)/sqrt(n)
            return(dft)
          }
```

We have shown above that $\{J_n(\omega_k)\}_k$ are close to uncorrelated and have variance close to $f(\omega_k)$. This means that the ratio $J_n(\omega_k)/f(\omega_k)^{1/2}$ are close to uncorrelated with variance close to one. Let us treat

$$Z_k = \frac{J_n(\omega_k)}{\sqrt{f(\omega_k)}},$$

as the transformed random variables, noting that $\{Z_k\}$ is complex, our aim is to show that the acf corresponding to $\{Z_k\}$ is close to zero. Of course, in practice we do not know the spectral density function f, therefore we estimate it using the piece of code (where `test` is the time series)

```
k<-kernel("daniell",6)
temp2 <-spec.pgram(test,k, taper=0, log = "no")$spec
n <- length(temp2)
temp3 <- c(temp2[c(1:n)],temp2[c(n:1)])
```

`temp3` simply takes a local average of the periodogram about the frequency of interest (however it is worth noting that `spec.pgram` does not do precisely this, which can be a bit annoying). In Section 11.3 we explain why this is a **consistent** estimator of the spectral density function. Notice that we also double the length, because the estimator `temp2` only gives estimates in the interval $[0, \pi]$. Thus our estimate of $\{Z_k\}$, which we denote as $\widehat{Z}_k = J_n(\omega_k)/\widehat{f}_n(\omega_k)^{1/2}$ is

```
temp1 <- dft(test);  temp4 <- temp1/sqrt(temp3)
```

We want to evaluate the covariance of $\{\widehat{Z}_k\}$ over various lags

$$\widehat{C}_n(r) = \frac{1}{n}\sum_{k=1}^{n} \widehat{Z}_k \overline{\widehat{Z}}_{k+r} = \frac{1}{n}\sum_{k=1}^{n} \frac{J_n(\omega_k)\overline{J_n(\omega_{k+r})}}{\sqrt{\widehat{f}_n(\omega_k)\widehat{f}_n(\omega_{k+r})}}$$

To speed up the evaluation, we use we can exploit the speed of the FFT, Fast Fourier Transform.

A plot of the AR(2) model

$$\varepsilon_t = 1.5\varepsilon_{t-1} - 0.75\varepsilon_{t-2} + \epsilon_t.$$

together with the real and imaginary parts of its DFT autocovariance is given in Figure 10.6. We observe that most of the correlations lie between $[-1.96, 1.96]$ (which corresponds to the 2.5% limits of a standard normal). Note that the 1.96 corresponds to the 2.5% limits, however this bound only holds if the time series is Gaussian. If the time series is non-Gaussian some corrections have to be made (see Dwivedi and Subba Rao (2011) and Jentsch and Subba Rao (2014)).

Exercise 10.1 *(a) Simulate an AR(2) process and run the above code using the sample size*

(i) $n = 64$ *(however use* `k<-kernel("daniell",3)`*)*

(ii) $n = 128$ *(however use* `k<-kernel("daniell",4)`*)*

Does the 'near decorrelation property' hold when the sample size is very small. Explain your answer by looking at the proof of the lemma.

(b) Simulate a piecewise stationary time series (this is a simple example of a nonstationary time series) by stringing two stationary time series together. One example is

```
ts1 = arima.sim(list(order=c(2,0,0), ar = c(1.5, -0.75)), n=128);
ts2 = arima.sim(list(order=c(1,0,0), ar = c(0.7)), n=128)
test = c(ts1/sd(ts1),ts2/sd(ts2))
```

Make a plot of this time series. Calculate the DFT covariance of this time series, what do you observe in comparison to the stationary case?

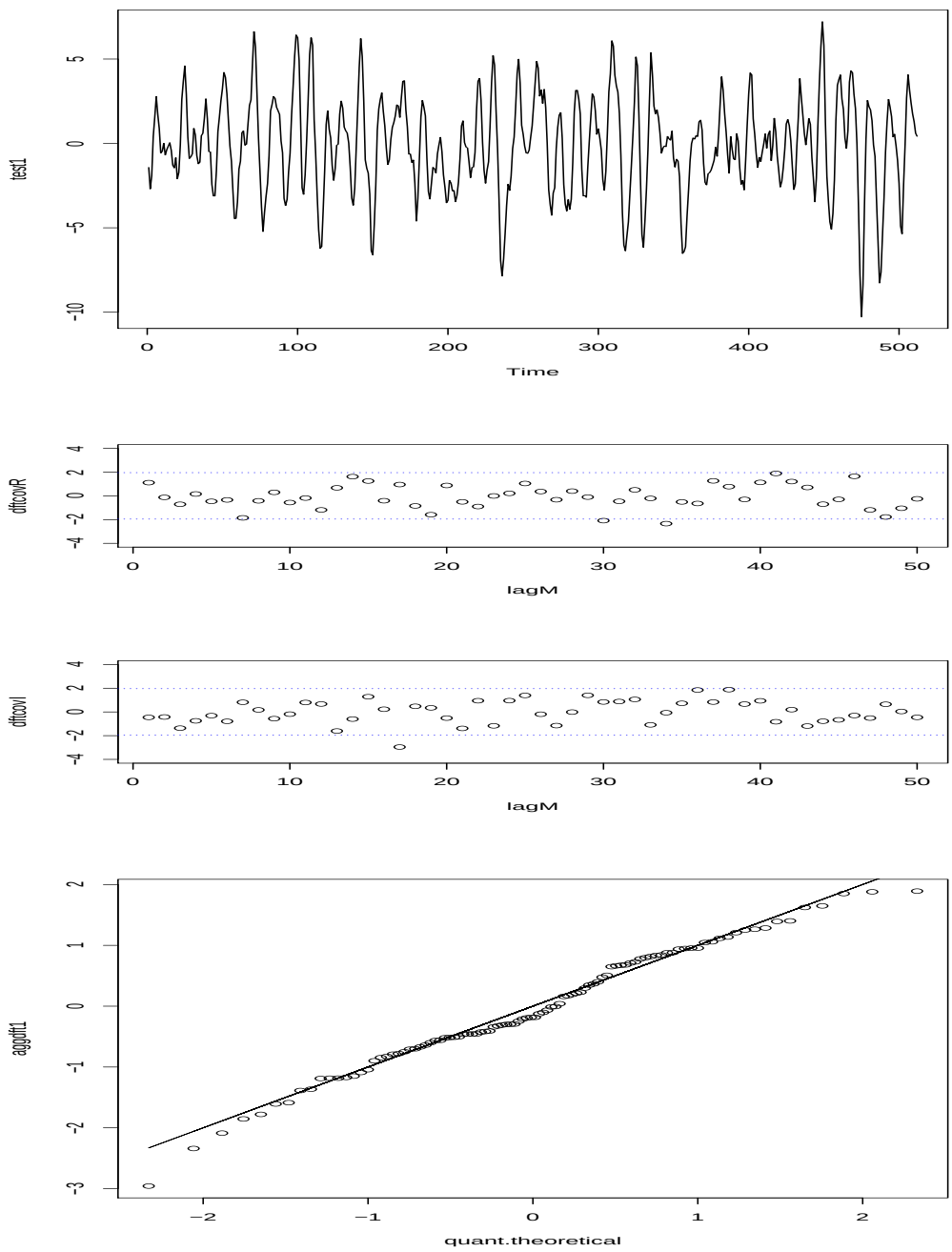


Figure 10.6: Top: Realization. Middle: Real and Imaginary of $\sqrt{n}\widehat{C}_n(r)$ plotted against the 'lag' r. Bottom: QQplot of the real and imaginary $\sqrt{n}\widehat{C}_n(r)$ against a standard normal.

10.2.2 Proof of Lemma 10.2.1

We will calculate $\text{cov}(J_n(\frac{2\pi k_1}{n}), J_n(\frac{2\pi k_2}{n}))$. The important aspect of this proof is that if we can isolate the exponentials, then we can use (10.1). It is this that gives rise to the near uncorrelatedness

property. Furthermore, since

$$\exp(i\frac{2\pi}{n}jk) = \exp(ij\omega_k) = \exp(ik\omega_j),$$

hence we can interchange the above terms. The proof hinges on the zeroing property of the DFT

$$\sum_{t=1}^{n} \exp\left(it\frac{2\pi j}{n}\right) = \begin{cases} 0 & j \notin n\mathbb{Z} \\ n & j \in n\mathbb{Z} \end{cases}.$$

It is possible to prove this result separating the complex variable into sines and cosines, but the proof is cumbersome. Instead, we summarize some well known properties of covariances of complex random variables. Suppose A is a complex random variables, then $\mathrm{var}[A]$ must be positive (not complex!), thus we define

$$\mathrm{var}[A] = \mathrm{E}\left[(A - \mathrm{E}[A])\overline{(A - \mathrm{E}[A])}\right] = \mathrm{E}[A\overline{A}] - |\mathrm{E}[A]|^2.$$

Based on this definition the covariance between two complex random variables is

$$\mathrm{cov}[A, B] = \mathrm{cov}\left[(A - \mathrm{E}[A])\overline{(B - \mathrm{E}[B])}\right] = \mathrm{E}[A\overline{B}] - \mathrm{E}[A]\overline{\mathrm{E}[B]}.$$

Note that $\mathrm{cov}[A, B] = \overline{\mathrm{cov}[B, A]}$.

Using this definitions we can write the covariance between two DFTs as

$$\mathrm{cov}\left(J_n(\frac{2\pi k_1}{n}), J_n(\frac{2\pi k_2}{n})\right) = \frac{1}{n}\sum_{t,\tau=1}^{n} \mathrm{cov}(X_t, X_\tau) \exp\left(i(tk_1 - \tau k_2)\frac{2\pi}{n}\right).$$

Next we change variables with $r = t - \tau$, this gives (for $0 \le k_1, k_2 < n$)

$$\begin{aligned} & \mathrm{cov}\left(J_n(\frac{2\pi k_1}{n}), J_n(\frac{2\pi k_2}{n})\right) \\ = \; & \frac{1}{n}\sum_{r=-(n-1)}^{n-1} c(r) \exp\left(-ir\frac{2\pi k_2}{n}\right) \sum_{t=1}^{n-|r|} \exp\left(\frac{2\pi it(k_1 - k_2)}{n})\right). \end{aligned}$$

Observe, if the limits of the inner sum were replaced with $\sum_{t=1}^{n-|r|}$, then we can use the zero property

of the DFTs. Thus in the next step we replace the limits of sum

$$\mathrm{cov}\left(J_n(\frac{2\pi k_1}{n}), J_n(\frac{2\pi k_2}{n})\right) = \sum_{r=-(n-1)}^{n-1} c(r)\exp\left(ir\frac{2\pi k_2}{n}\right)\underbrace{\frac{1}{n}\sum_{t=1}^{n}\exp\left(\frac{2\pi it(k_1-k_2)}{n}\right)}_{\delta_{k_1}(k_2)} - R_n.$$

The remainder is what is additional term added to sum and is

$$R_n = \frac{1}{n}\sum_{r=-(n-1)}^{n-1} c(r)\exp\left(-ir\frac{2\pi k_2}{n}\right)\sum_{t=n-|r|+1}^{n}\exp\left(\frac{2\pi it(k_1-k_2)}{n})\right)$$

Thus $|R_n| \leq \frac{1}{n}\sum_{|r|\leq n}|rc(r)| = O(n^{-1})$. Therefore

$$\begin{aligned}&\mathrm{cov}\left(J_n(\frac{2\pi k_1}{n}), J_n(\frac{2\pi k_2}{n})\right)\\ =\ &\frac{1}{n}\sum_{r=-(n-1)}^{n-1} c(r)\exp\left(-ir\frac{2\pi k_2}{n}\right)\sum_{t=1}^{n-|r|}\exp\left(\frac{2\pi it(k_1-k_2)}{n})\right).\end{aligned}$$

Observe, if the limits of the inner sum were replaced with $\sum_{t=1}^{n}$, then we can use the zero property of the DFTs. Thus in the next step we replace the limits of sum

$$\begin{aligned}\mathrm{cov}\left(J_n(\frac{2\pi k_1}{n}), J_n(\frac{2\pi k_2}{n})\right) &= \sum_{r=-(n-1)}^{n-1} c(r)\exp\left(ir\frac{2\pi k_2}{n}\right)\delta_{k_1}(k_2) - R_n\\ &= f_n(\omega_{k_1})\delta_{k_1}(k_2) + O(n^{-1}),\end{aligned}$$

thus proving the result. □

10.2.3 The DFT and complete decorrelation

The proof above is very "hands on". But it does not adequately answer the question of where the $O(n^{-1})$ actually comes from. Nor how it can be removed. It transpires that by using simple ideas from linear prediction (as described in Chapters 6 and 7), one can obtain a deep understanding on the role of the DFT in the analysis of stationary time series. These insights allow us to connect time domain estimation methods to the frequency domain estimation (described in Chapter 11). The ideas presented here are based on work done jointly with Junho Yang (see Subba Rao and Yang (2020)).

As in the previous section, the derivations are based following simple identity on the special

zeroing property of sums of DFTs:

$$\frac{1}{n}\sum_{t=1}^{n}\exp\left(it\omega_{k_1-k_2}\right) = \begin{cases} 0 & k_1-k_2 \notin n\mathbb{Z} \\ 1 & k_1-k_2 \in n\mathbb{Z} \end{cases}.$$

Let us return to the very simple calculations in the previous section. In particular, the product of the DFTs

$$\begin{aligned} J_n\left(\frac{2\pi k_1}{n}\right)\overline{J}_n\left(\frac{2\pi k_2}{n}\right) &= \frac{1}{n}\sum_{t,\tau=1}^{n} X_t X_\tau e^{it\omega_{k_1}-i\tau\omega_{k_2}} \\ &= \frac{1}{n}\sum_{t=1}^{n} e^{it(\omega_{k_1}-\omega_{k_2})} X_t \sum_{\tau=1}^{n} X_\tau e^{i(t-\tau)\omega_{k_2}}. \end{aligned} \tag{10.7}$$

Thus the covariance between the DFTs is

$$\operatorname{cov}\left[J_n(\frac{2\pi k_1}{n}), J_n(\frac{2\pi k_2}{n})\right] = \frac{1}{n}\sum_{t=1}^{n} e^{it(\omega_{k_1}-\omega_{k_2})}\sum_{\tau=1}^{n} c(t-\tau)e^{i(t-\tau)\omega_{k_2}}.$$

If the inner sum of the above were replaced with $\sum_{\tau=-\infty}^{\infty} c(t-\tau)e^{i(t-\tau)\omega_{k_2}} = f(\omega_{k_2})$, then we immediately have the decorrelation property; this is because the inner sum would not depend on t and the inner and outer sums can be separated out.

A graphic for the inner summand in the term above is given Figure 10.7. The terms show what has been missed. The omission of these terms give rise to the $O(n^{-1})$ error. Furthermore, we observe that when $t = 1$, there is large error as the term $n^{-1}\sum_{\tau=-\infty}^{0} c(1-\tau)e^{i(1-\tau)\omega_k}$ has been omitted. However, if t is more central and far from the boundaries, then only the far left and right tails are omitted and the resulting error term is small. To remove the error term, we need a method for correcting for the red terms $c(t-\tau)$ for all τ outside the observed domain of observation $[1, 2, \ldots, n-1, n]$. Of course, X_τ is unknown outside the interval of observation. Nevertheless, one can still construct variables which yield $c(t-\tau)$. To find these we review what we have learnt from linear prediction.

We return to some results in Section 5.1.5. Define the space X, where $X = \operatorname{sp}(X_1, \ldots, X_n)$ and $P_X(Y)$ denote the linear projection of the random variable Y onto X. Then equation (5.6) in Section 5.1.5 shows

$$\operatorname{cov}\left(P_X(Y), X_\ell\right) = \operatorname{cov}(Y, X_\ell) \qquad 1 \leq \ell \leq n.$$

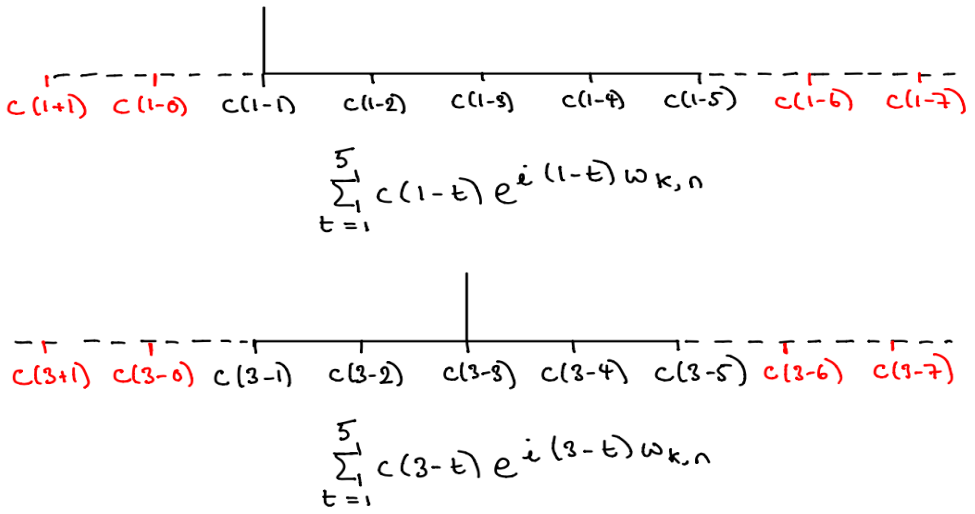


Figure 10.7: An illustration of the correlation between two DFTs (with $n = 5$). The red terms are terms that have been omitted and contribute to the $O(n^{-1})$ error.

The above is a general result, we now apply it to the above problem. Suppose $\{X_t\}$ be a second order stationary time series and set $X = \text{sp}(X_1, \ldots, X_n)$, these are the random variables in the domain of observation. We define the random variables outside the domain of observation and let $Y = X_\tau$ for $\tau \neq \{1, \ldots, n\}$. By using the above result we have

$$\text{cov}\left(P_X(X_\tau), X_t\right) = \text{cov}(X_\tau, X_t) = c(\tau - t) \qquad 1 \leq t \leq n. \tag{10.8}$$

Now $P_X(X_\tau) \in \text{sp}(X_1, \ldots, X_n)$, this tells us we can correct the boundary by including random variables that belong the domain of observation.

Worked Example: AR(1) Suppose that $X_t = \phi X_{t-1} + \varepsilon_t$ where $|\phi| < 1$. We have learn from Section 6.2.2 (on partial covariance of a time series) that second order stationarity of a time series implies that the coefficients for forecasting into the future and forecasting into the past are the same. Further, we have shown in Section 7.2 that prediction/forecasting for AR(p) models are simply the AR(p) coefficients. Using these two facts the best linear predictor of X_τ given $\{X_t\}_{t=1}^n$ is

$$\begin{aligned} P_X(X_{n+1}) &= \phi X_n \text{ and } P_X(X_{n+r}) = \phi^r X_n \text{ for } r \geq 1 \\ P_X(X_0) &= \phi X_1 \text{ and } P_X(X_{1-r}) = \phi^{|r|} X_1 \text{ for } r \geq 1. \end{aligned} \tag{10.9}$$

See Figure 10.8 for a graphic of the above.

Figure 10.8

It is straightfoward to verify (10.8) for the AR(1) model. We have shown in Section 6.1.2 for $r \geq 1$ that

$$\text{cov}\left(X_n, X_{n+r}\right) = \frac{\phi^{|r|}}{1-\phi^2}.$$

On the other hand using the predictor in (10.9) we have

$$\text{cov}\left(X_n, P_X(X_{n+r})\right) = \phi^{|r|}\text{var}[X_n] = \frac{\phi^{|r|}}{1-\phi^2}.$$

And we observe the two expression match. □

The above calculations show $P_X(X_\tau)$ is a "proxy" for X_τ, that is only in terms of the observed data, but also contains all the required information on X_τ. Thus turning to (10.7) we replace *one* of the DFTs

$$\frac{1}{n}\sum_{t=1}^{n} e^{it(\omega_{k_1}-\omega_{k_2})} X_t \sum_{\tau=1}^{n} X_\tau e^{i(t-\tau)\omega_{k_2}}$$

with

$$\begin{aligned} & \frac{1}{n}\sum_{t=1}^{n} e^{it(\omega_{k_1}-\omega_{k_2})} X_t \sum_{\tau=1}^{n} X_\tau e^{i(t-\tau)\omega_{k_2}} + \frac{1}{n}\sum_{t=1}^{n} e^{it(\omega_{k_1}-\omega_{k_2})} X_t \sum_{\tau\neq\{1,\ldots,n\}} P_X(X_\tau) e^{i(t-\tau)\omega_{k_2}} \\ = & \frac{1}{n}\sum_{t=1}^{n} e^{it(\omega_{k_1}-\omega_{k_2})} X_t \sum_{\tau=-\infty}^{\infty} P_X(X_\tau) e^{i(t-\tau)\omega_{k_2}}, \end{aligned}$$

where we note that the last line of the above is the best linear predictor $P_X(X_\tau) = X_\tau$ if $\tau \in \{1, \ldots, n\}$. The above can be written as a product:

$$\frac{1}{n}\sum_{t=1}^{n} e^{it(\omega_{k_1}-\omega_{k_2})} X_t \sum_{\tau=-\infty}^{\infty} P_X(X_\tau) e^{i(t-\tau)\omega_{k_2}} = \left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t e^{it\omega_{k_1}}\right)\left(\frac{1}{\sqrt{n}}\sum_{\tau=-\infty}^{\infty} P_X(X_\tau) e^{-i\tau\omega_{k_1}}\right).$$

The first term on the right hand side of the above is the regular DFT. The second term is a variant of the regular DFT which includes the regular DFT but linearly predicts outside the boundary (we call this the complete DFT):

$$\widetilde{J}_n(\omega;f) = J_n(\omega) + \frac{1}{\sqrt{n}} \sum_{\tau \neq \{1,2,\ldots,n\}} P_X(X_\tau) e^{i\tau\omega} = \frac{1}{\sqrt{n}} \sum_{\tau=-\infty}^{\infty} P_X(X_\tau) e^{i\tau\omega}.$$

The calculations above show that

$$\begin{aligned} \mathrm{cov}[J_n(\omega_{k_1}), \widetilde{J}_n(\omega_{k_2};f)] &= \frac{1}{n}\sum_{t=1}^{n} e^{it(\omega_{k_1}-\omega_{k_2})} f(\omega_{k_2}) \\ &= f(\omega_{k_2}) \frac{1}{n}\sum_{t=1}^{n} e^{it(\omega_{k_1}-\omega_{k_2})} = f(\omega_{k_2})\delta_{k_1}(k_2). \end{aligned}$$

We summarize the result in the following Lemma.

Lemma 10.2.2 *Suppose $\{X_t\}$ is a second order stationary time series, where $\sum_r |rc(r)| < \infty$ and $f = \sum_{\in\mathbb{Z}} c(r)\exp(ir\omega)$. Then we have*

$$\mathrm{cov}\left[J_n\left(\frac{2\pi k_1}{n}\right), J_n\left(\frac{2\pi k_2}{n};f\right)\right] = \begin{cases} f(\frac{2\pi k}{n}) & k_1 = k_2 \\ 0 & 1 \leq k_1 \neq k_2 \leq n/2 \end{cases}.$$

Worked Example: AR(1) For the AR(1) model, it can easily be shown that

$$\widetilde{J}_n(\omega;f) = J_n(\omega) + \frac{\phi}{\sqrt{n}}\left(\frac{1}{\phi(\omega)}X_1 + \frac{e^{i(n+1)\omega}}{\overline{\phi(\omega)}}X_n\right),$$

where $\phi(\omega) = 1 - \phi e^{-i\omega}$. We observe that the predictive contribution is actually rather small (if n is large). □

As the contribution of the prediction term in the complete DFT is actually quite small $O(n^{-1/2})$. It is this that allows us to focus the time series analysis on just the regular DFT (with the cost of a $O(n^{-1})$ bias).

10.3 Summary of spectral representation results

In this section we summarize some spectral properties. We do this by considering the DFT of the data $\{J_n(\omega_k)\}_{k=1}^n$. It is worth noting that to calculate $\{J_n(\omega_k)\}_{k=1}^n$ is computationally very fast and requires only $O(n \log n)$ computing operations (see Section A.5, where the Fast Fourier Transform is described).

10.3.1 The spectral (Cramer's) representation theorem

We observe that for any sequence $\{X_t\}_{t=1}^n$ that it can be written as the inverse transform for $1 \leq t \leq n$

$$X_t = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} J_n(\omega_k) \exp(-it\omega_k), \tag{10.10}$$

which can be written as an integral

$$X_t = \sum_{k=2}^{n} \exp(-it\omega_k) \left[Z_n(\omega_k) - Z_n(\omega_{k-1}) \right] = \int_0^{2\pi} \exp(-it\omega) dZ_n(\omega), \tag{10.11}$$

where $Z_n(\omega) = \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor \frac{\omega}{2\pi} n \rfloor} J_n(\omega_k)$.

The second order stationary property of X_t means that the DFT $J_n(\omega_k)$ is close to an uncorrelated sequence or equivalently the process $Z_n(\omega)$ has near 'orthogonal' increments, meaning that for any two non-intersecting intervals $[\omega_1, \omega_2]$ and $[\omega_3, \omega_4]$ that $Z_n(\omega_2) - Z_n(\omega_1)$ and $Z_n(\omega_4) - Z_n(\omega_3)$. The spectral representation theorem generalizes this result, it states that for any second order stationary time series $\{X_t\}$ there exists an a process $\{Z(\omega); \omega \in [0, 2\pi]\}$ where for all $t \in \mathbb{Z}$

$$X_t = \int_0^{2\pi} \exp(-it\omega) dZ(\omega) \tag{10.12}$$

and $Z(\omega)$ has orthogonal increments, meaning that for any two non-intersecting intervals $[\omega_1, \omega_2]$ and $[\omega_3, \omega_4]$ $\mathrm{E}[Z(\omega_2) - Z(\omega_1)][Z(\omega_2) - Z(\omega_1)] = 0$.

We now explore the relationship between the DFT with the orthogonal increment process.

Using (10.12) we see that

$$\begin{aligned} J_n(\omega_k) &= \frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n} X_t \exp(it\omega_k) = \frac{1}{\sqrt{2\pi n}}\int_0^{2\pi}\left(\sum_{t=1}^{n}\exp(it[\omega_k-\omega])\right) dZ(\omega) \\ &= \frac{1}{\sqrt{2\pi n}}\int_0^{2\pi}\left(e^{i(n+1)(\omega_k-\omega_0)/2}D_{n/2}(\omega_k-\omega)\right) dZ(\omega), \end{aligned}$$

where $D_{n/2}(x) = \sin[((n+1)/2)x]/\sin(x/2)$ is the Dirichlet kernel (see Priestley (1983), page 419). We recall that the Dirichlet kernel limits to the Dirac-delta function, therefore very crudely speaking we observe that the DFT is an approximation of the orthogonal increment localized about ω_k (though mathematically this is not strictly correct).

10.3.2 Bochner's theorem

This is a closely related result that is stated in terms of the so called spectral distribution. First the heuristics. We see that from Lemma 10.2.1 that the DFT $J_n(\omega_k)$, is close to uncorrelated. Using this and inverse Fourier transforms we see that for $1 \leq t, \tau \leq n$ we have

$$\begin{aligned} c(t-\tau) = \mathrm{cov}(X_t, X_\tau) &= \frac{1}{n}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\mathrm{cov}\left(J_n(\omega_{k_1}), J_n(\omega_{k_2})\right)\exp(-it\omega_{k_1}+i\tau\omega_{k_2}) \\ &\approx \frac{1}{n}\sum_{k=1}^{n}\mathrm{var}(J_n(\omega_k))\exp(-i(t-\tau)\omega_k). \end{aligned} \tag{10.13}$$

Let $F_n(\omega) = \frac{1}{n}\sum_{k=1}^{\lfloor \frac{\omega}{2\pi}n\rfloor}\mathrm{var}[J_n(\omega_k)]$, then the above can be written as

$$c(t-\tau) \approx \int_0^{2\pi}\exp(-i(t-\tau)\omega)dF_n(\omega),$$

where we observe that $F_n(\omega)$ is a positive function which in non-decreasing over ω. Bochner's theorem is an extension of this is states that for any autocovariance function $\{c(k)\}$ we have the representation

$$c(t-\tau) = \int_0^{2\pi}\exp(-i(t-\tau)\omega)f(\omega)d\omega = \int_0^{2\pi}\exp(-i(t-\tau)\omega)dF(\omega).$$

where $F(\omega)$ is a positive non-decreasing bounded function. Moreover, $F(\omega) = \mathrm{E}(|Z(\omega)|^2)$. We note that if the spectral density function exists (which is only true if $\sum_r |c(r)|^2 < \infty$) then $F(\omega) = \int_0^\omega f(\lambda)d\lambda$.

Remark 10.3.1 *The above results hold for both linear and nonlinear time series, however, in the case that X_t has a linear representation*

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

then X_t has the particular form

$$X_t = \int A(\omega) \exp(-ik\omega) dZ(\omega), \tag{10.14}$$

where $A(\omega) = \sum_{j=-\infty}^{\infty} \psi_j \exp(ij\omega)$ and $Z(\omega)$ is an orthogonal increment process, but in addition $\mathrm{E}(|dZ(\omega)|^2) = d\omega$ ie. the variance of increments do not vary over frequency (as this varying has been absorbed by $A(\omega)$, since $F(\omega) = |A(\omega)|^2$).

We mention that a more detailed discussion on spectral analysis in time series is give in Priestley (1983), Chapters 4 and 6, Brockwell and Davis (1998), Chapters 4 and 10, Fuller (1995), Chapter 3, Shumway and Stoffer (2006), Chapter 4. In many of these references they also discuss tests for periodicity etc (see also Quinn and Hannan (2001) for estimation of frequencies etc.).

10.4 The spectral density and spectral distribution

10.4.1 The spectral density and some of its properties

Finally, having made ourselves familiar with the DFT and the spectral density function we can prove Theorem 3.4.2, which relates the autocovariance with the positiveness of its Fourier transform. In the following lemma we consider absolutely summable autocovariances, in a later theorem (called Bochner's theorem) we show that any valid autocovariance has this representation.

Theorem 10.4.1 (Positiveness of the spectral density) *Suppose the coefficients $\{c(k)\}$ 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. Moreover

$$c(k) = \int_0^{2\pi} \exp(-ik\omega) f(\omega) d\omega. \tag{10.15}$$

It is worth noting that f is called the spectral density corresponding to the covariances $\{c(k)\}$.

PROOF. We first show that if $\{c(k)\}$ is a non-negative definite sequence, then $f(\omega)$ is a nonnegative function. We recall that since $\{c(k)\}$ is non-negative then for any sequence $\underline{x} = (x_1, \ldots, x_N)$ (real or complex) we have $\sum_{s,t=1}^{n} x_s c(s-t)\bar{x}_s \geq 0$ (where $\bar{x}_s$ is the complex conjugate of x_s). Now we consider the above for the particular case $\underline{x} = (\exp(i\omega), \ldots, \exp(in\omega))$. Define the function

$$f_n(\omega) = \frac{1}{2\pi n} \sum_{s,t=1}^{n} \exp(is\omega) c(s-t) \exp(-it\omega).$$

Thus by definition $f_n(\omega) \geq 0$. We note that $f_n(\omega)$ can be rewritten as

$$f_n(\omega) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{(n-1)} \left(\frac{n-|k|}{n} \right) c(k) \exp(ik\omega).$$

Comparing $f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} c(k) \exp(ik\omega)$ with $f_n(\omega)$ we see that

$$\begin{aligned} \big| f(\omega) - f_n(\omega) \quad &\leq \quad \frac{1}{2\pi} \big| \sum_{|k| \geq n} c(k) \exp(ik\omega) \big| + \frac{1}{2\pi} \big| \sum_{k=-(n-1)}^{(n-1)} \frac{|k|}{n} c(k) \exp(ik\omega) \big| \\ &:= \quad I_n + II_n. \end{aligned}$$

Since $\sum_{k=-\infty}^{\infty} |c(k)| < \infty$ it is clear that $I_n \to 0$ as $n \to \infty$. Using Lemma A.1.1 we have $II_n \to 0$ as $n \to \infty$. Altogether the above implies

$$\big| f(\omega) - f_n(\omega) \big| \to 0 \quad \text{as } n \to \infty. \tag{10.16}$$

Now it is clear that since for all n, $f_n(\omega)$ are nonnegative functions, the limit f must be nonnegative (if we suppose the contrary, then there must exist a sequence of functions $\{f_{n_k}(\omega)\}$ which are not necessarily nonnegative, which is not true). Therefore we have shown that if $\{c(k)\}$ is a nonnegative definite sequence, then $f(\omega)$ is a nonnegative function.

We now show the converse, that is the Fourier coefficients of any non-negative ℓ_2 function $f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} c(k) \exp(ik\omega)$, is a positive semi-definite sequence. Writing $c(k) = \int_0^{2\pi} f(\omega) \exp(ik\omega) d\omega$

we substitute this into Definition 3.4.1 to give

$$\sum_{s,t=1}^{n} x_s c(s-t)\bar{x}_s = \int_0^{2\pi} f(\omega)\Big\{\sum_{s,t=1}^{n} x_s \exp(i(s-t)\omega)\bar{x}_s\Big\}d\omega = \int_0^{2\pi} f(\omega)\left|\sum_{s=1}^{n} x_s \exp(is\omega)\right|^2 d\omega \geq 0.$$

Hence we obtain the desired result. □

The above theorem is very useful. It basically gives a simple way to check whether a sequence $\{c(k)\}$ is non-negative definite or not (hence whether it is a covariance function - recall Theorem 3.4.1). See Brockwell and Davis (1998), Corollary 4.3.2 or Fuller (1995), Theorem 3.1.9, for alternative explanations.

Example 10.4.1 *Consider the empirical covariances (here we gives an alternative proof to Remark 8.2.1) defined in Chapter 8*

$$\hat{c}_n(k) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-|k|} X_t X_{t+|k|} & |k| \leq n-1 \\ 0 & \textit{otherwise} \end{cases},$$

we give an alternative proof to Lemma 8.2.1 to show that $\{\hat{c}_n(k)\}$ is non-negative definite sequence. To show that the sequence we take the Fourier transform of $\hat{c}_n(k)$ and use Theorem 10.4.1. The Fourier transform of $\{\hat{c}_n(k)\}$ is

$$\sum_{k=-(n-1)}^{(n-1)} \exp(ik\omega)\hat{c}_n(k) = \sum_{k=-(n-1)}^{(n-1)} \exp(ik\omega)\frac{1}{n}\sum_{t=1}^{n-|k|} X_t X_{t+|k|} = \frac{1}{n}\Big|\sum_{t=1}^{n} X_t \exp(it\omega)\Big| \geq 0.$$

Since the above is non-negative, this means that $\{\hat{c}_n(k)\}$ is a non-negative definite sequence.

We now state a useful result which relates the largest and smallest eigenvalue of the variance of a stationary process to the smallest and largest values of the spectral density (we recall we used this in Lemma 7.14.2).

Lemma 10.4.1 *Suppose that $\{X_k\}$ is a stationary process with covariance function $\{c(k)\}$ and spectral density $f(\omega)$. Let $\Sigma_n = \mathrm{var}(\underline{X}_n)$, where $\underline{X}_n = (X_1, \ldots, X_n)$. Suppose $\inf_\omega f(\omega) \geq m > 0$ and $\sup_\omega f(\omega) \leq M < \infty$ Then for all n we have*

$$\lambda_{\min}(\Sigma_n) \geq \inf_\omega f(\omega) \quad \textit{and} \quad \lambda_{\max}(\Sigma_n) \leq \sup_\omega f(\omega).$$

PROOF. Let $\underline{e}_1$ be the eigenvector with smallest eigenvalue λ_1 corresponding to Σ_n. Then using $c(s-t) = \int f(\omega)\exp(i(s-t)\omega)d\omega$ we have

$$\begin{aligned}\lambda_{\min}(\Sigma_n) &= \underline{e}_1'\Sigma_n\underline{e}_1 = \sum_{s,t=1}^{n}\bar{e}_{s,1}c(s-t)e_{t,1} = \int f(\omega)\sum_{s,t=1}^{n}\bar{e}_{s,1}\exp(i(s-t)\omega)e_{t,1}d\omega = \\ &= \int_0^{2\pi} f(\omega)\left|\sum_{s=1}^{n}e_{s,1}\exp(is\omega)\right|^2 d\omega \geq \inf_{\omega} f(\omega)\int_0^{2\pi}\left|\sum_{s=1}^{n}e_{s,1}\exp(is\omega)\right|^2 d\omega = \inf_{\omega} f(\omega),\end{aligned}$$

since by definition $\int|\sum_{s=1}^{n}e_{s,1}\exp(is\omega)|^2 d\omega = \sum_{s=1}^{n}|e_{s,1}|^2 = 1$ (using Parseval's identity). Using a similar method we can show that $\lambda_{\max}(\Sigma_n) \leq \sup f(\omega)$. □

We now state a version of the above result which requires weaker conditions on the autocovariance function (only that they decay to zero).

Lemma 10.4.2 *Suppose the covariance $\{c(k)\}$ decays to zero as $k \to \infty$, then for all n, $\Sigma_n = \mathrm{var}(\underline{X}_n)$ is a non-singular matrix (Note we do not require the stronger condition the covariances are absolutely summable).*

PROOF. See Brockwell and Davis (1998), Proposition 5.1.1. □

10.4.2 The spectral distribution and Bochner's (Hergoltz) theorem

Theorem 10.4.1 hinges on the result that $f_n(\omega) = \sum_{r=-(n-1)}^{(n-1)}(1-|r|/n)e^{ir\omega}$ has a well defined pointwise limit as $n \to \infty$, this only holds when the sequence $\{c(k)\}$ is absolutely summable. Of course this may not always be the case. An extreme example is the time series $X_t = Z$. Clearly this is a stationary time series and its covariance is $c(k) = \mathrm{var}(Z) = 1$ for all k. In this case the autocovariance sequence $\{c(k) = 1; k \in \mathbb{Z}\}$, is not absolutely summable, hence the representation of the covariance in Theorem 10.4.1 does not apply. The reason is because the Fourier transform of the infinite sequence $\{c(k) = 1; k \in \mathbb{Z}\}$ is not well defined (clearly $\{c(k) = 1\}_k$ does not belong to ℓ_1).

However, we now show that Theorem 10.4.1 can be generalised to include <u>all</u> non-negative definite sequences and stationary processes, by considering the spectral distribution rather than the spectral density.

Theorem 10.4.2 *A function $\{c(k)\}$ is non-negative definite sequence if and only if*

$$c(k) = \int_0^{2\pi} \exp(-ik\omega) dF(\omega), \tag{10.17}$$

where $F(\omega)$ is a right-continuous (this means that $F(x+h) \to F(x)$ as $0 < h \to 0$), non-decreasing, non-negative, bounded function on $[-\pi, \pi]$ (hence it has all the properties of a distribution and it can be consider as a distribution - it is usually called the spectral distribution). This representation is unique.

This is a very constructive result. It shows that the Fourier coefficients of any distribution function form a non-negative definite sequence, and thus, if $c(k) = c(-k)$ (hence is symmetric) correspond to the covarance function of a random process. In Figure 10.9 we give two distribution functions. the top plot is continuous and smooth, therefore it's derivative will exist, be positive and belong to ℓ_2. So it is clear that its Fourier coefficients form a non-negative definite sequence. The interesting aspect of Thereom 10.4.2 is that the Fourier coefficients corresponding to the distribution function in the second plot also forms a non-negative definite sequence even though the derivative of this distribution function does not exist. However, this sequence will not belong to ℓ_2 (ie. the correlations function will not decay to zero as the lag grows).

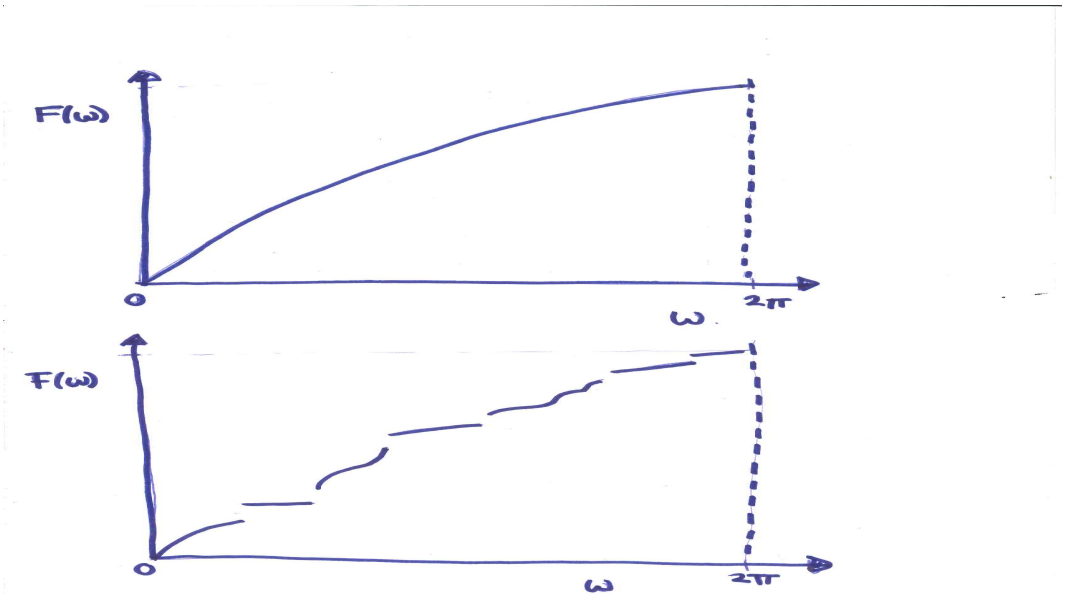


Figure 10.9: Both plots are of non-decreasing functions, hence are valid distribution functions. The top plot is continuous and smooth, thus its derivative (the spectral density function) exists. Whereas the bottom plot is not (spectral density does not exist).

Example 10.4.2 *Using the above we can construct the spectral distribution for the (rather silly) time series $X_t = Z$. Let $F(\omega) = 0$ for $\omega < 0$ and $F(\omega) = \mathrm{var}(Z)$ for $\omega \geq 0$ (hence F is the step function). Then we have*

$$\mathrm{cov}(X_t, X_{t+k}) = \mathrm{var}(Z) = \int \exp(-ik\omega) dF(\omega).$$

Example 10.4.3 *Consider the second order stationary time series*

$$X_t = U_1 \cos(\lambda t) + U_2 \sin(\lambda t),$$

where U_1 and U_2 are iid random variables with mean zero and variance σ^2 and λ the frequency. It can be shown that

$$\mathrm{cov}(X_t, X_{t+k}) = \frac{\sigma^2}{2} \left[\exp(i\lambda k) + \exp(-i\lambda k)\right].$$

Observe that this covariance does not decay with the lag k. Then

$$\mathrm{cov}(X_t, X_{t+k}) = \mathrm{var}(Z) = \int_0^{2\pi} \exp(-ik\omega) dF(\omega).$$

where

$$F(\omega) = \begin{cases} 0 & \omega < -\lambda \\ \sigma^2/2 & -\lambda \leq \omega < \lambda \\ \sigma^2 & \lambda \geq \omega. \end{cases}$$

10.5 The spectral representation theorem

We now state the spectral representation theorem and give a rough outline of the proof.

Theorem 10.5.1 *If $\{X_t\}$ is a second order stationary time series with mean zero, and spectral distribution $F(\omega)$, and the spectral distribution function is $F(\omega)$, then there exists a right continuous, orthogonal increment process $\{Z(\omega)\}$ (that is $\mathrm{E}[(Z(\omega_1) - Z(\omega_2))\overline{(Z(\omega_3) - Z(\omega_4))}] = 0$, when the intervals $[\omega_1, \omega_2]$ and $[\omega_3, \omega_4]$ do not overlap) such that*

$$X_t = \int_0^{2\pi} \exp(-it\omega) dZ(\omega), \tag{10.18}$$

where for $\omega_1 \geq \omega_2$, $\mathrm{E}|Z(\omega_1) - Z(\omega_2)|^2 = F(\omega_1) - F(\omega_2)$ (noting that $F(0) = 0$). (One example of a right continuous, orthogonal increment process is Brownian motion, though this is just one example, and usually $Z(\omega)$ will be far more general than Brownian motion).

Heuristically we see that (10.18) is the decomposition of X_t in terms of frequencies, whose amplitudes are orthogonal. In other words X_t is decomposed in terms of frequencies $\exp(it\omega)$ which have the orthogonal amplitudes $dZ(\omega) \approx (Z(\omega + \delta) - Z(\omega))$.

Remark 10.5.1 *Note that so far we have not defined the integral on the right hand side of (10.18). It is known as a stochastic integral. Unlike many deterministic functions (functions whose derivative exists), one cannot really suppose $dZ(\omega) \approx Z'(\omega)d\omega$, because usually a typical realisation of $Z(\omega)$ will not be smooth enough to differentiate. For example, it is well known that Brownian is quite 'rough', that is a typical realisation of Brownian motion satisfies $|B(t_1, \bar{\omega}) - B(t_2, \bar{\omega})| \leq K(\bar{\omega})|t_1 - t_t|^{\gamma}$, where $\bar{\omega}$ is a realisation and $\gamma \leq 1/2$, but in general γ will not be larger. The integral $\int g(\omega)dZ(\omega)$ is well defined if it is defined as the limit (in the mean squared sense) of discrete sums. More precisely, let $Z_n(\omega) = \sum_{k=1}^{n} Z(\omega_k)I_{\omega_{n_k-1},\omega_{n_k}}(\omega) = \sum_{k=1}^{\lfloor n\omega/2\pi \rfloor}[Z(\omega_k) - Z(\omega_{k-1})]$, then*

$$\int g(\omega)dZ_n(\omega) = \sum_{k=1}^{n} g(\omega_k)\{Z(\omega_k) - Z(\omega_{k-1})\}.$$

The limit of $\int g(\omega)dZ_n(\omega)$ as $n \to \infty$ is $\int g(\omega)dZ(\omega)$ (in the mean squared sense, that is $\mathrm{E}[\int g(\omega)dZ(\omega) - \int g(\omega)dZ_n(\omega)]^2$). Compare this with our heuristics in equation (10.11).

For a more precise explanation, see Parzen (1959), Priestley (1983), Sections 3.6.3 and Section 4.11, page 254, and Brockwell and Davis (1998), Section 4.7. For a very good review of elementary stochastic calculus see Mikosch (1999).

A very elegant explanation on the different proofs of the spectral representation theorem is given in Priestley (1983), Section 4.11. We now give a rough outline of the proof using the functional theory approach.

Remark 10.5.2 *We mention that the above representation applies to both stationary and nonstationary time series. What makes the exponential functions $\{\exp(ik\omega)\}$ special is if a process is stationary then the representation of $c(k) := \mathrm{cov}(X_t, X_{t+k})$ in terms of exponentials is guaranteed:*

$$c(k) = \int_0^{2\pi} \exp(-ik\omega)dF(\omega). \tag{10.19}$$

Therefore there always exists an orthogonal random function $\{Z(\omega)\}$ such that

$$X_t = \int \exp(-it\omega) dZ(\omega).$$

Indeed, whenever the exponential basis is used in the definition of either the covariance or the process $\{X_t\}$, the resulting process will always be second order stationary.

Brockwell and Davis (1998), Proposition 4.8.2 states an interesting consequence of the spectral representation theorem. Suppose that $\{X_t\}$ is a second order stationary time series with spectral distribution $F(\omega)$. If $F(\omega)$ has a discontinuity at λ_0, then X_t almost surely has the representation

$$X_t = \int_0^{2\pi} e^{it\omega} dZ(\omega) + e^{it\lambda_0} \left(Z(\lambda_0^+) - Z(\lambda_0^-) \right)$$

where $Z(\lambda_0^-)$ and $Z(\lambda_0^+)$ denote the left and right limit. This result means that discontinuities in the spectral distribution mean that the corresponding time series contains a deterministic sinusoid functions i.e.

$$X_t = A\cos(\lambda_0 t) + B\sin(\lambda_0 t) + \varepsilon_t$$

where ε_t is a stationary time series. We came across this "feature" in Section 2.5. If the spectral distribution contains a discontinuity, then "formally" the spectral density (which is the derivative of the spectral distribution) is the dirac-delta function at the discontinuity. The periodogram is a "crude" (inconsistent) estimator of the spectral density function, however it captures the general features of the underlying spectral density. Look at Figures 2.11-2.13, observe that there is a large peak corresponding the deterministic frequency and that this peak grows taller as the sample size n grows. This large peak is limiting to the dirac delta function.

Finally we state Brockwell and Davis (1998), Proposition 4.9.1, which justifies our use of the DFT. Brockwell and Davis (1998), Proposition 4.9.1 states that if $\{X_t\}$ is a second order stationary time series with spectral distribution F and ν_1 and ν_2 are continuity points of F then

$$\frac{1}{2\pi} \sum_{|t|\leq n} X_t \int_{\nu_1}^{\nu_2} \exp(it\omega) d\omega \to Z(\nu_2) - Z(\nu_1),$$

where the convergence is in mean squared.

Let $\omega_k = 2\pi k/n$, then using this result we have

$$\frac{1}{2\pi\sqrt{n}} \sum_{|t|\leq n} X_t \exp(it\omega_k) \approx \sqrt{n} \sum_{|t|\leq n} X_t \int_{\omega_k}^{\omega_{k+1}} \exp(it\omega) d\omega \approx \sqrt{n} \left[Z(\omega_{k+1}) - Z(\omega_k) \right],$$

without the scaling factor $\sqrt{n}$, the above would limit to zero. Thus as claimed previously, the DFT estimates the "increments".

10.6 The spectral density functions of MA, AR and ARMA models

We obtain the spectral density function for MA(∞) processes. Using this we can easily obtain the spectral density for ARMA processes. Let us suppose that $\{X_t\}$ satisfies the representation

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j} \tag{10.20}$$

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2 and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. We recall that the covariance of above is

$$c(k) = \mathrm{E}(X_t X_{t+k}) = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+k}. \tag{10.21}$$

Since $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, it can be seen that

$$\sum_k |c(k)| \leq \sum_k \sum_{j=-\infty}^{\infty} |\psi_j| \cdot |\psi_{j+k}| < \infty.$$

Hence by using Theorem 10.4.1, the spectral density function of $\{X_t\}$ is well defined. There are several ways to derive the spectral density of $\{X_t\}$, we can either use (10.21) and $f(\omega) = \frac{1}{2\pi} \sum_k c(k) \exp(ik\omega)$ or obtain the spectral representation of $\{X_t\}$ and derive $f(\omega)$ from the spectral representation. We prove the results using the latter method.

10.6.1 The spectral representation of linear processes

Since $\{\varepsilon_t\}$ are iid random variables, using Theorem 10.5.1 there exists an orthogonal random function $\{Z(\omega)\}$ such that

$$\varepsilon_t = \int_0^{2\pi} \exp(-it\omega) dZ_\varepsilon(\omega).$$

Since $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{E}(\varepsilon_t^2) = \sigma^2$ multiplying the above by ε_t, taking expectations and noting that due to the orthogonality of $\{Z_\varepsilon(\omega)\}$ we have $\mathrm{E}(dZ_\varepsilon(\omega_1)\overline{dZ_\varepsilon(\omega_2)}) = 0$ unless $\omega_1 = \omega_2$ we have that $\mathrm{E}(|dZ_\varepsilon(\omega)|^2) = \sigma^2 d\omega$, hence $f_\varepsilon(\omega) = (2\pi)^{-1}\sigma^2$.

Using the above we obtain the following spectral representation for $\{X_t\}$, where X_t is a linear time series

$$X_t = \sum_{j=\infty}^{\infty} \psi_j \varepsilon_{t-j} \int_0^{2\pi} \left\{ \sum_{j=-\infty}^{\infty} \psi_j \exp(ij\omega) \right\} \exp(-it\omega) dZ_\varepsilon(\omega).$$

Hence

$$X_t = \int_0^{2\pi} A(\omega) \exp(-it\omega) dZ_\varepsilon(\omega) = \int_0^{2\pi} \exp(-it\omega) dZ_X(\omega) \tag{10.22}$$

where $A(\omega) = \sum_{j=-\infty}^{\infty} \psi_j \exp(ij\omega)$ and $Z_X(\omega) = A(\omega) Z_\varepsilon(\omega)$. We note that this is the unique spectral representation of X_t.

Definition 10.6.1 (The Cramer Representation) *We mention that the representation in (10.22) of a stationary process is usually called the Cramer representation of a stationary process, where*

$$X_t = \int_0^{2\pi} A(\omega) \exp(-it\omega) dZ(\omega),$$

where $\{Z(\omega) : 0 \leq \omega \leq 2\pi\}$ are orthogonal functions.

Exercise 10.2 *(i) Suppose that $\{X_t\}$ has an $MA(1)$ representation $X_t = \theta\varepsilon_t + \varepsilon_{t-1}$. What is its Cramer's representation?*

(ii) Suppose that $\{X_t\}$ has a causal $AR(1)$ representation $X_t = \phi X_{t-1} + \varepsilon_t$. What is its Cramer's representation?

10.6.2 The spectral density of a linear process

Multiplying (10.22) by X_{t+k} and taking expectations gives

$$\mathrm{E}(X_tX_{t+k}) = c(k) = \int_0^{2\pi} A(\omega_1)A(-\omega_2)\exp(-i(t+k)\omega_1 + it\omega_2)\mathrm{E}(dZ(\omega_1)d\overline{Z(\omega_2)}).$$

Due to the orthogonality of $\{Z(\omega)\}$ we have $\mathrm{E}(dZ(\omega_1)d\overline{Z(\omega_2)}) = 0$ unless $\omega_1 = \omega_2$, altogether this gives

$$\mathrm{E}(X_tX_{t+k}) = c(k) = \int_0^{2\pi} |A(\omega)|^2\exp(-ik\omega)\mathrm{E}(|dZ(\omega)|^2) = \int_0^{2\pi} f(\omega)\exp(-ik\omega)d\omega,$$

where $f(\omega) = \frac{\sigma^2}{2\pi}|A(\omega)|^2$. Comparing the above with (10.15) we see that $f(\cdot)$ is the spectral density function.

The spectral density function corresponding to the linear process defined in (10.20) is

$$f(\omega) = \frac{\sigma^2}{2\pi}|\sum_{j=-\infty}^{\infty}\psi_j\exp(-ij\omega)|^2.$$

Remark 10.6.1 (An alternative, more hands on proof) *An alternative proof which avoids the Cramer representation is to use that the acf of a linear time series is $c(r) = \sigma^2\sum_k \psi_j\psi_{j+r}$ (see Lemma 6.1.1). Thus by definition the spectral density function is*

$$\begin{aligned} f(\omega) &= \frac{1}{2\pi}\sum_{r=-\infty}^{\infty} c(r)\exp(ir\omega) \\ &= \frac{\sigma^2}{2\pi}\sum_{r=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}\psi_j\psi_{j+r}\exp(ir\omega). \end{aligned}$$

Now make a change of variables $s = j + r$ this gives

$$f(\omega) = \frac{\sigma^2}{2\pi}\sum_{s=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}\psi_j\psi_s\exp(i(s-j)\omega) = \frac{\sigma^2}{2\pi}\left|\sum_{j=-\infty}^{\infty}\psi_j e^{ij\omega}\right|^2 = \frac{\sigma^2}{2\pi}|A(\omega)|^2.$$

Example 10.6.1 *Let us suppose that $\{X_t\}$ is a stationary $ARMA(p,q)$ time series (not necessarily invertible or causal), where*

$$X_t - \sum_{j=1}^{p}\psi_j X_{t-j} = \sum_{j=1}^{q}\theta_j\varepsilon_{t-j},$$

$\{\varepsilon_t\}$ are iid random variables with $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{E}(\varepsilon_t^2) = \sigma^2$. Then the spectral density of $\{X_t\}$ is

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{|1 + \sum_{j=1}^{q} \theta_j \exp(ij\omega)|^2}{|1 - \sum_{j=1}^{q} \phi_j \exp(ij\omega)|^2}$$

We note that because the ARMA is the ratio of trignometric polynomials, this is known as a rational spectral density.

Remark 10.6.2 *The roots of the characteristic function of an AR process will have an influence on the location of peaks in its corresponding spectral density function. To see why consider the $AR(2)$ model*

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

where $\{\varepsilon_t\}$ are iid random variables with zero mean and $\mathrm{E}(\varepsilon^2) = \sigma^2$. Suppose the roots of the characteristic polynomial $\phi(B) = 1 - \phi_1 B - \phi_2 B^2$ lie outside the unit circle and are complex conjugates where $\lambda_1 = r\exp(i\theta)$ and $\lambda_2 = r\exp(-i\theta)$. Then the spectral density function is

$$\begin{aligned} f(\omega) &= \frac{\sigma^2}{|1 - r\exp(i(\theta - \omega))|^2 |1 - r\exp(i(-\theta - \omega)|^2} \\ &= \frac{\sigma^2}{[1 + r^2 - 2r\cos(\theta - \omega)][1 + r^2 - 2r\cos(\theta + \omega)]}. \end{aligned}$$

If $r > 0$, the $f(\omega)$ is maximum when $\omega = \theta$, on the other hand if, $r < 0$ then the above is maximum when $\omega = \theta - \pi$. Thus the peaks in $f(\omega)$ correspond to peaks in the pseudo periodicities of the time series and covariance structure (which one would expect), see Section 6.1.2. How pronounced these peaks are depend on how close r is to one. The close r is to one the larger the peak. We can generalise the above argument to higher order Autoregressive models, in this case there may be multiple peaks. In fact, this suggests that the larger the number of peaks, the higher the order of the AR model that should be fitted.

10.6.3 Approximations of the spectral density to AR and MA spectral densities

In this section we show that the spectral density

$$f(\omega) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} c(r) \exp(ir\omega)$$

can be approximated to any order by the spectral density of an AR(p) or MA(q) process.

We do this by truncating the infinite number of covariances by a finite number, however, this does not necessarily lead to a positive definite spectral density. This can easily be proven by noting that

$$\widetilde{f}_m(\omega) = \sum_{r=-m}^{m} c(r) \exp(ir\omega) = \int_0^{2\pi} f(\lambda) D_m(\omega - \lambda) d\lambda,$$

where $D_m(\lambda) = \sin[(n + 1/2)\lambda]/\sin(\lambda/2)$. Observe that $D_m(\cdot)$ can be negative, which means that $\widetilde{f}_m(\omega)$ can be negative despite f being positive.

Example 10.6.2 *Consider the AR(1) process* $X_t = 0.75X_{t-1} + \varepsilon_t$ *where* $\text{var}[\varepsilon_t] = 1$. *In Lemma 6.1.1 we showed that the autcovariance corresponding to this model is* $c(r) = [1 - 0.75^2]^{-1} 0.75^{|r|}$.

Let us define a process whose autocorrelation is $\tilde{c}(0) = [1 - 0.75^2]^{-1}$, $c(1) = c(-1) = [1 - 0.75^2]^{-1} 0.75$ *and* $\tilde{c}(r) = 0$ *for* $|r| > 1$. *The 'spectral density' of this process is*

$$\widetilde{f}_m(\omega) = \frac{1}{1 - 0.75^2} \left(1 + 2 \times \frac{3}{4} \cos[\omega] \right).$$

It is clear that this function can be zero for some values of ω. *This means that* $\{\tilde{c}(r)\}$ *is not a well defined covariance function, hence there does not exist a time series with this covariance structure. In other words, simply truncating an autocovariance is not enough to guarantee that it positive definite sequence.*

Instead we consider a slight variant on this and define

$$\frac{1}{2\pi} \sum_{r=-m}^{m} \left(1 - \frac{|r|}{m} \right) c(r) \exp(ir\omega)$$

which is positive.

Remark 10.6.3 *We note that f_m is known as a Cesáro sum because it can be written as*

$$f_m(\omega) = \frac{1}{2\pi}\sum_{r=-m}^{m}\left(1-\frac{|r|}{m}\right)c(r)\exp(ir\omega) = \frac{1}{m}\sum_{n=0}^{m}\widetilde{f}_n(\omega), \tag{10.23}$$

where $\widetilde{f}_n(\cdot) = \frac{1}{2\pi}\sum_{r=-n}^{n} c(r)\exp(ir\omega)$. Strangely, there is no guarantee that the truncated Fourier transform $\widetilde{f}_n$ is not negative, however $f_n(\cdot)$ is definitely positive. There are are a few ways to prove this:

(i) The first method we came across previously, $\mathrm{var}[J_n(\omega)] = f_n(\omega)$, it is clear that using this construction $\inf_\omega f_n(\omega) \geq 0$.

(ii) By using (10.23) we can write $f_m(\cdot)$ as

$$f_m(\omega) = \int_0^{2\pi} f(\lambda)F_m(\omega-\lambda)d\lambda,$$

where $F_m(\lambda) = \frac{1}{m}\sum_{r=-m}^{m} D_r(\lambda) = \frac{1}{m}\left(\frac{\sin(n\lambda/2)}{\sin(\lambda/2)}\right)^2$ and $D_r(\lambda) = \sum_{j=-r}^{r}\exp(ij\omega)$ (these are the Fejer and Dirichlet kernels respectively). Since both f and F_m are positive, then f_m has to be positive.

The Cesaro sum is special in the sense that

$$\sup_\omega |f_m(\omega) - f(\omega)| \to 0, \qquad \text{as } m\to\infty. \tag{10.24}$$

Thus for a large enough m, $f_m(\omega)$ will be within δ of the spectral density f. Using this we can prove the results below.

Lemma 10.6.1 *Suppose that $\sum_r |c(r)| < \infty$, f is the spectral density of the covariances and $\inf_{\omega\in[0,2\pi]} f(\omega) > 0$. Then for every $\delta > 0$, there exists a m such that $|f(\omega) - f_m(\omega)| < \delta$ and $f_m(\omega) = \sigma^2|\psi(\omega)|^2$, where $\psi(\omega) = \sum_{j=0}^{m}\psi_j\exp(ij\omega)$. Thus we can approximate the spectral density of f with the spectral density of a MA.*

PROOF. We show that there exists an MA(m) which has the spectral density $f_m(\omega)$, where f_m is defined in (10.23). Thus by (10.24) we have the result.

Before proving the result we note that if a "polynomial" is of the form

$$p(z) = a_0 + \sum_{j=1}^{m} a_j\left(z+z^{-1}\right)$$

then it has the factorization $p(z) = C\prod_{j=1}^{m}[1-\lambda_j z][1-\lambda_j^{-1}z]$, where λ_j is such that $|\lambda_j| < 1$. Furthermore, if $\{a_j\}_{j=0}^{m}$ are real and $z^m p(z)$ has no roots on the unit circle, then the coefficients of the polynomial $\prod_{j=1}^{m}[1-\lambda_j z]$ are real. The above claims are true because

(i) To prove that $p(z) = C\prod_{j=1}^{m}[1-\lambda_j z][1-\lambda_j^{-1}z]$, we note that $z^m p(z)$ is a $2m$-order polynomial. Thus it can be factorized. If there exists a root λ whose inverse is not a root, then the resulting polynomial will have not have the symmetric structure.

(ii) By the complex conjugate theorem, since $z^m p(z)$ has real coefficients, then its complex roots must be conjugates. Moreover, since no roots lie on the unit circle, then no conjugates lie on the unit circle. Thus the coefficients of $\prod_{j=1}^{m}[1-\lambda_j z]$ are real (if it did lie on the unit circle, then we can distribute the two roots between the two polynomials).

Thus setting $z = e^{i\omega}$

$$\sum_{r=-m}^{m} a_r \exp(ir\omega) \;=\; C\prod_{j=1}^{m}[1-\lambda_j\exp(i\omega)]\left[1-\lambda_j^{-1}\exp(-i\omega)\right].$$

for some finite constant C. We use the above result. Since $\inf f_m(\omega) > 0$ and setting $a_r = [1-|r|n^{-1}]c(r)$, we can write f_m as

$$\begin{aligned} f_m(\omega) &= K\left[\prod_{j=1}^{m}(1-\lambda_j^{-1}\exp(i\omega))\right]\left[\prod_{j=1}^{m}(1-\lambda_j\exp(-i\omega))\right] \\ &= A(\omega)A(-\omega) = |A(\omega)|^2, \end{aligned}$$

where

$$A(z) = \prod_{j=1}^{m}(1-\lambda_j^{-1}z).$$

Since $A(z)$ is an mth order polynomial where all the roots are greater than 1, we can always construct an MA(m) process which has $A(z)$ as its 'transfer' function. Thus there exists an MA(m) process which has $f_m(\omega)$ as its spectral density function. □

Remark 10.6.4 *(i) The above result requires that* $\inf_\omega f(\omega) > 0$*, in order to ensure that* $f_m(\omega)$ *is strictly positive. This assumption can be relaxed (and the proof becomes a little more complicated), see Brockwell and Davis (1998), Theorem 4.4.3.*

(ii)

Lemma 10.6.2 *Suppose that $\sum_r |c(r)| < \infty$ and f is corresponding the spectral density function where $\inf_\omega f(\omega) > 0$. Then for every $\delta > 0$, there exists a m such that $|f(\omega) - g_m(\omega)| < \delta$ and $g_m(\omega) = \sigma^2|\phi(\omega)^{-1}|^2$, where $\phi(\omega) = \sum_{j=0}^m \phi_j \exp(ij\omega)$ and the roots of $\phi(z)$ lie outside the unit circle. Thus we can approximate the spectral density of f with the spectral density of a causal autoregressive process.*

PROOF. We first note that we can write

$$\big|f(\omega) - g_m(\omega)\big| \quad = \quad f(\omega)|g_m(\omega)^{-1} - f(\omega)^{-1}|g_m(\omega).$$

Since $f(\cdot) \in L_2$ and is bounded away from zero, then $f^{-1} \in L_2$ and we can write f^{-1} as

$$f^{-1}(\omega) = \sum_{r=\infty}^{\infty} d_r \exp(ir\omega),$$

where d_r are the Fourier coefficients of f^{-1}. Since f is positive and symmetric, then f^{-1} is positive and symmetric such that $f^{-1}(\omega) = \sum_{r=-\infty}^{\infty} d_r e^{ir\omega}$ and $\{d_r\}$ is a positive definite symmetric sequence. Thus we can define the positive function g_m where

$$g_m^{-1}(\omega) = \sum_{|r|\leq m} \left(1 - \frac{|r|}{m}\right) d_r \exp(ir\omega)$$

and is such that $|g_m^{-1}(\omega) - f^{-1}(\omega)| < \delta$, which implies

$$\big|f(\omega) - g_m(\omega)\big| \leq [\sum_r |c(r)|]^2\delta.$$

Now we can apply the same arguments to prove to Lemma 10.6.1 we can show that g_m^{-1} can be factorised as $g_m^{-1}(\omega) = C|\phi_m(\omega)|^2$ (where ϕ_m is an mth order polynomial whose roots lie outside the unit circle). Thus $g_m(\omega) = C|\phi_m(\omega)|^{-2}$ and we obtain the desired result. □

10.7 Cumulants and higher order spectrums

We recall that the covariance is a measure of linear dependence between two random variables. Higher order cumulants are a measure of higher order dependence. For example, the third order

cumulant for the zero mean random variables X_1, X_2, X_3 is

$$\text{cum}(X_1, X_2, X_3) = \text{E}(X_1 X_2 X_3)$$

and the fourth order cumulant for the zero mean random variables X_1, X_2, X_3, X_4 is

$$\text{cum}(X_1, X_2, X_3, X_4) = \text{E}(X_1 X_2 X_3 X_4) - \text{E}(X_1 X_2)\text{E}(X_3 X_4) - \text{E}(X_1 X_3)\text{E}(X_2 X_4) - \text{E}(X_1 X_4)\text{E}(X_2 X_3).$$

From the definition we see that if X_1, X_2, X_3, X_4 are independent then $\text{cum}(X_1, X_2, X_3) = 0$ and $\text{cum}(X_1, X_2, X_3, X_4) = 0$.

Moreover, if X_1, X_2, X_3, X_4 are Gaussian random variables then $\text{cum}(X_1, X_2, X_3) = 0$ and $\text{cum}(X_1, X_2, X_3, X_4) = 0$. Indeed all cumulants higher than order two is zero. This comes from the fact that cumulants are the coefficients of the power series expansion of the logarithm of the characteristic function of $\{X_t\}$, which is

$$g_X(t) = i \underbrace{\underline{\mu}'}_{\text{mean}} \underline{t} - \frac{1}{2} \underline{t}' \underbrace{\Sigma}_{\text{cumulant}} \underline{t}.$$

Since the spectral density is the Fourier transform of the covariance it is natural to ask whether one can define the higher order spectral density as the fourier transform of the higher order cumulants. This turns out to be the case, and the higher order spectra have several interesting properties. Let us suppose that $\{X_t\}$ is a stationary time series (notice that we are assuming it is strictly stationary and not second order). Let $\kappa_3(t,s) = \text{cum}(X_0, X_t, X_s)$, $\kappa_3(t,s,r) = \text{cum}(X_0, X_t, X_s, X_r)$ and $\kappa_q(t_1, \ldots, t_{q-1}) = cum(X_0, X_{t_1}, \ldots, X_{t_q})$ (noting that like the covariance the higher order cumulants are invariant to shift). The third, fourth and the general qth order spectras is defined as

$$\begin{aligned}
f_3(\omega_1, \omega_2) &= \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} \kappa_3(s,t) \exp(is\omega_1 + it\omega_2) \\
f_4(\omega_1, \omega_2, \omega_3) &= \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} \kappa_4(s,t,r) \exp(is\omega_1 + it\omega_2 + ir\omega_3) \\
f_q(\omega_1, \omega_2, \ldots, \omega_{q-1}) &= \sum_{t_1, \ldots, t_{q-1}=-\infty}^{\infty} \kappa_q(t_1, t_2, \ldots, t_{q-1}) \exp(it_1\omega_1 + it_2\omega_2 + \ldots + it_{q-1}\omega_{q-1}).
\end{aligned}$$

Example 10.7.1 (Third and Fourth order spectral density of a linear process) *Let us sup-*

pose that $\{X_t\}$ *satisfies*

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

where $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, $\mathrm{E}(\varepsilon_t) = 0$ *and* $\mathrm{E}(\varepsilon_t^4) < \infty$. *Let* $A(\omega) = \sum_{j=-\infty}^{\infty} \psi_j \exp(ij\omega)$. *Then it is straightforward to show that*

$$\begin{aligned} f(\omega) &= \sigma^2 |A(\omega)|^2 \\ f_3(\omega_1, \omega_2) &= \kappa_3 A(\omega_1) A(\omega_2) A(-\omega_1 - \omega_2) \\ f_4(\omega_1, \omega_2, \omega_3) &= \kappa_4 A(\omega_1) A(\omega_2) A(\omega_3) A(-\omega_1 - \omega_2 - \omega_3), \end{aligned}$$

where $\kappa_3 = \mathrm{cum}(\varepsilon_t, \varepsilon_t, \varepsilon_t)$ *and* $\kappa_4 = \mathrm{cum}(\varepsilon_t, \varepsilon_t, \varepsilon_t, \varepsilon_t)$.

We see from the example, that unlike the spectral density, the higher order spectras are not necessarily positive or even real.

A review of higher order spectra can be found in Brillinger (2001). Higher order spectras have several applications especially in nonlinear processes, see Subba Rao and Gabr (1984). We will consider one such application in a later chapter.

Using the definition of the higher order spectrum we can now generalise Lemma 10.2.1 to higher order cumulants (see Brillinger (2001), Theorem 4.3.4).

Proposition 10.7.1 $\{X_t\}$ *is a strictly stationary time series, where for all* $1 \leq i \leq q-1$ *we have* $\sum_{t_1,\ldots,t_{q-1}=\infty}^{\infty} |(1+t_i)\kappa_q(t_1, \ldots, t_{q-1})| < \infty$ *(note that this is simply a generalization of the covariance assumption* $\sum_r |rc(r)| < \infty$*). Then we have*

$$\begin{aligned} \mathrm{cum}(J_n(\omega_{k_1}), \ldots, J_n(\omega_{k_q})) &= \frac{1}{n^{q/2}} f_q(\omega_{k_2}, \ldots, \omega_{k_q}) \sum_{j=1}^{n} \exp(ij(\omega_{k_1} - \ldots - \omega_{k_q})) + O(\frac{1}{n^{q/2}}) \\ &= \begin{cases} \frac{1}{n^{(q-1)/2}} f_q(\omega_{k_2}, \ldots, \omega_{k_q}) + O(\frac{1}{n^{q/2}}) & \sum_{i=1}^{q} k_i = n\mathbb{Z} \\ O(\frac{1}{n^{q/2}}) & \textit{otherwise} \end{cases} \end{aligned}$$

where $\omega_{k_i} = \frac{2\pi k_i}{n}$.

10.8 Extensions

10.8.1 The spectral density of a time series with randomly missing observations

Let us suppose that $\{X_t\}$ is a second order stationary time series. However $\{X_t\}$ is not observed at everytime point and there are observations missing, thus we only observe X_t at $\{\tau_k\}_k$. Thus what is observed is $\{X_{\tau_k}\}$. The question is how to deal with this type of data. One method was suggested in **?**. He suggested that the missingness mechanism $\{\tau_k\}$ be modelled stochastically. That is define the random process $\{Y_t\}$ which only takes the values $\{0,1\}$, where $Y_t = 1$ if X_t is observed, but $Y_t = 0$ if X_t is not observed. Thus we observe $\{X_tY_t\}_t = \{X_{t_k}\}$ and also $\{Y_t\}$ (which is the time points the process is observed). He also suggests modelling $\{Y_t\}$ as a stationary process, which is independent of $\{X_t\}$ (thus the missingness mechanism and the time series are independent).

The spectral densities of $\{X_tY_t\}$, $\{X_t\}$ and $\{Y_t\}$ have an interest relationship, which can be exploited to estimate the spectral density of $\{X_t\}$ given estimators of the spectral densities of $\{X_tY_t\}$ and $\{X_t\}$ (which we recall are observed). We first note that since $\{X_t\}$ and $\{Y_t\}$ are stationary, then $\{X_tY_t\}$ is stationary, furthermore

$$\begin{aligned}\mathrm{cov}(X_tY_t, X_\tau Y_\tau) &= \mathrm{cov}(X_t, X_\tau)\mathrm{cov}(Y_t, Y_\tau) + \mathrm{cov}(X_t, Y_\tau)\mathrm{cov}(Y_t, X_\tau) + cum(X_t, Y_t, X_\tau, Y_\tau) \\ &= \mathrm{cov}(X_t, X_\tau)\mathrm{cov}(Y_t, Y_\tau) = c_X(t-\tau)c_Y(t-\tau)\end{aligned}$$

where the above is due to independence of $\{X_t\}$ and $\{Y_t\}$. Thus the spectral density of $\{X_tY_t\}$ is

$$\begin{aligned}f_{XY}(\omega) &= \frac{1}{2\pi}\sum_{r=-\infty}^{\infty}\mathrm{cov}(X_0Y_0, X_rY_r)\exp(ir\omega) \\ &= \frac{1}{2\pi}\sum_{r=-\infty}^{\infty}c_X(r)c_Y(r)\exp(ir\omega) \\ &= \int f_X(\lambda)f_Y(\omega-\lambda)d\omega,\end{aligned}$$

where $f_X(\lambda) = \frac{1}{2\pi}\sum_{r=-\infty}^{\infty}c_X(r)\exp(ir\omega)$ and $f_Y(\lambda) = \frac{1}{2\pi}\sum_{r=-\infty}^{\infty}c_Y(r)\exp(ir\omega)$ are the spectral densities of the observations and the missing process.

10.9 Appendix: Some proofs

PROOF of Theorem 10.4.2. We first show that if $\{c(k)\}$ is non-negative definite sequence, then we can write $c(k) = \int_0^{2\pi} \exp(ik\omega) dF(\omega)$, where $F(\omega)$ is a distribution function.

To prove the result we adapt some of the ideas used to prove Theorem 10.4.1. As in the proof of Theorem 10.4.1 define the (nonnegative) function

$$f_n(\omega) = \text{var}[J_n(\omega)] = \frac{1}{2\pi n} \sum_{s,t=1}^{n} \exp(is\omega) c(s-t) \exp(-it\omega) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{(n-1)} \left(\frac{n-|k|}{n} \right) c(k) \exp(ik\omega).$$

If $\{c(k)\}$ is not absolutely summable, the limit of $f_n(\omega)$ is no longer well defined. Instead we consider its integral, which will always be a distribution function (in the sense that it is nondecreasing and bounded). Let us define the function $F_n(\omega)$ whose derivative is $f_n(\omega)$, that is

$$F_n(\omega) = \int_0^{\omega} f_n(\lambda) d\lambda = \frac{\omega}{2\pi} c(0) + \frac{2}{2\pi} \sum_{r=1}^{n-1} \left(1 - \frac{r}{n}\right) c(r) \frac{\sin(\omega r)}{r} \quad 0 \leq \lambda \leq 2\pi.$$

Since $f_n(\lambda)$ is nonnegative, $F_n(\omega)$ is a nondecreasing function. Furthermore it is bounded since

$$F_n(2\pi) = \int_0^{2\pi} f_n(\lambda) d\lambda = c(0).$$

Hence F_n satisfies all properties of a distribution and can be treated as a distribution function. This means that we can apply Helly's theorem to the sequence $\{F_n\}_n$. We first recall that if $\{x_n\}$ are real numbers defined on a compact set $X \subset \mathbb{R}$, then there exists a subsequence $\{x_{n_m}\}_m$ which has a limit in the set X (this is called the Bolzano-Weierstrass theorem). An analogous result exists for measures, this is called Helly's theorem (see Ash (1972), page 329). It states that for any sequence of distributions $\{G_n\}$ defined on $[0, 2\pi]$, were $G_n(0) = 0$ and $\sup_n G_n(2\pi) < M < \infty$, there exists a subsequence $\{n_m\}_m$ where $G_{n_m}(x) \to G(x)$ as $m \to \infty$ for each $x \in [0, 2\pi]$ at which G is continuous. Furthermore, since $G_{n_m}(x) \to G(x)$ (pointwise as $m \to \infty$), this implies (see Varadhan, Theorem 4.1 for equivalent forms of convergence) that for any bounded sequence h we have that

$$\int h(x) dG_{n_m}(x) \to \int h(x) dG(x) \qquad \text{as } m \to \infty.$$

We now apply this result to $\{F_n\}_n$. Using Helly's theorem there exists a subsequence of distributions

$\{F_{n_m}\}_m$ which has a pointwise limit F. Thus for any bounded function h we have

$$\int h(x)dF_{n_m}(x) \to \int h(x)dF(x) \qquad \text{as } m \to \infty. \tag{10.25}$$

We focus on the function $h(x) = \exp(-ik\omega)$. It is clear that for every k and n we have

$$\int_0^{2\pi} \exp(-ik\omega)dF_n(\omega) = \int_0^{2\pi} \exp(ik\omega)f_n(\omega)d\omega = \begin{cases} (1 - \frac{|k|}{n})c(k) & |k| \leq n \\ 0 & |k| \geq n \end{cases} \tag{10.26}$$

Define the sequence

$$d_{n,k} = \int_0^{2\pi} \exp(ik\omega)dF_n(\omega) = \left(1 - \frac{|k|}{n}\right) c(k).$$

We observe that for fixed k, $\{d_{n,k}; n \in \mathbb{Z}\}$ is a Cauchy sequence, where

$$d_{n,k} \to d_k = c(k) \tag{10.27}$$

as $n \to \infty$.

Now we use (10.25) and focus on the convergent subsequence $\{n_m\}_m$. By using (10.25) we have

$$d_{n_m,k} = \int \exp(-ikx)dF_{n_m}(x) \to \int \exp(-ikx)dF(x) \qquad \text{as } m \to \infty$$

and by (10.27) $d_{n_m,k} \to c(k)$ as $m \to \infty$. Thus

$$c(k) = \int \exp(-ikx)dF(x).$$

This gives the first part of the assertion.

To show the converse, that is $\{c(k)\}$ is a non-negative definite sequence when $c(k)$ is defined as $c(k) = \int \exp(ik\omega)dF(\omega)$, we use the same method given in the proof of Theorem 10.4.1, that is

$$\begin{aligned} \sum_{s,t=1}^{n} x_s c(s-t)\bar{x}_s &= \int_0^{2\pi} \Big\{ \sum_{s,t=1}^{n} x_s \exp(-i(s-t)\omega)\bar{x}_s \Big\} dF(\omega) \\ &= \int_0^{2\pi} \left| \sum_{s=1}^{n} x_s \exp(-is\omega) \right|^2 dF(\omega) \geq 0, \end{aligned}$$

since $F(\omega)$ is a distribution.

Finally, if $\{c(k)\}$ were absolutely summable, then we can use Theorem 10.4.1 to write $c(k) = \int_0^{2\pi} \exp(-ik\omega) dF(\omega)$, where $F(\omega) = \int_0^{\omega} f(\lambda) d\lambda$ and $f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} c(k) \exp(ik\omega)$. By using Theorem 10.4.1 we know that $f(\lambda)$ is nonnegative, hence $F(\omega)$ is a distribution, and we have the result. □

Rough PROOF of the Spectral Representation Theorem To prove the result we first define two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, where $\mathcal{H}_1$ one contains deterministic functions and $\mathcal{H}_2$ contains random variables.

First we define the space

$$\mathcal{H}_1 = \overline{\text{sp}}\{e^{it\omega}; t \in \mathbb{Z}\}$$

with inner-product

$$\langle f, g \rangle = \int_0^{2\pi} f(x) \overline{g(x)} dF(x) \tag{10.28}$$

(and of course distance $\langle f-g, f-g \rangle = \int_0^{2\pi} |f(x) - g(x)|^2 dF(x)$) it is clear that this inner product is well defined because $\langle f, f \rangle \geq 0$ (since F is a measure). It can be shown (see Brockwell and Davis (1998), page 144) that $\mathcal{H}_1 = \left\{ g; \int_0^{2\pi} |g(\omega)|^2 dF(\omega) < \infty \right\}$ [1]. We also define the space

$$\mathcal{H}_2 = \overline{\text{sp}}\{X_t; t \in \mathbb{Z}\}$$

with inner-product $\text{cov}(X, Y) = \text{E}[X\overline{Y}] - \text{E}[X]\text{E}[\overline{Y}]$.

Now let us define the linear mapping $T : \mathcal{H}_1 \to \mathcal{H}_2$

$$T(\sum_{j=1}^{n} a_j \exp(ik\omega)) = \sum_{j=1}^{n} a_j X_k, \tag{10.29}$$

for any n (it is necessary to show that this can be extended to infinite n, but we won't do so here). We will shown that T defines an isomorphism (ie. it is a one-to-one linear mapping that preserves norm). To show that it is a one-to-one mapping see Brockwell and Davis (1998), Section 4.7. It is clear that it is linear, there all that remains is to show that the mapping preserves inner-product.

[1] Roughly speaking it is because all continuous functions on $[0, 2\pi]$ are dense in $L_2([0, 2\pi], \mathcal{B}, F)$ (using the metric $\|f - g\| = \langle f - g, f - g \rangle$ and the limit of Cauchy sequences). Since all continuous function can be written as linear combinations of the Fourier basis, this gives the result.

Suppose $f, g \in \mathcal{H}_1$, then there exists coefficients $\{f_j\}$ and $\{g_j\}$ such that $f(x) = \sum_j f_j \exp(ij\omega)$ and $g(x) = \sum_j g_j \exp(ij\omega)$. Hence by definition of T in (10.29) we have

$$\langle Tf, Tg\rangle = \operatorname{cov}(\sum_j f_j X_j, \sum_j g_j X_j) = \sum_{j_1,j_2} f_{j_1}\overline{g_{j_2}}\operatorname{cov}(X_{j_1}, X_{j_2}) \tag{10.30}$$

Now by using Bochner's theorem (see Theorem 10.4.2) we have

$$\langle Tf, Tg\rangle = \int_0^{2\pi} \big(\sum_{j_1,j_2} f_{j_1}\overline{g_{j_2}} \exp(i(j_1 - j_2)\omega)\big) dF(\omega) = \int_0^{2\pi} f(x)\overline{g(x)} dF(x) = \langle f, g\rangle. \tag{10.31}$$

Hence $< Tf, Tg >=< f, g >$, so the inner product is preserved (hence T is an isometry).

Altogether this means that T defines an isomorphism betwen $\mathcal{H}_1$ and $\mathcal{H}_2$. Therefore all functions which are in $\mathcal{H}_1$ have a corresponding random variable in $\mathcal{H}_2$ which has similar properties.

For all $\omega \in [0, 2\pi]$, it is clear that the identity functions $I_{[0,\omega]}(x) \in \mathcal{H}_1$. Thus we define the random function $\{Z(\omega); 0 \leq \omega \leq 2\pi\}$, where $T(I_{[0,\omega]}(\cdot)) = Z(\omega) \in \mathcal{H}_2$ (since T is an isomorphism). Since that mapping T is linear we observe that

$$T(I_{[\omega_1,\omega_2]}) = T(I_{[0,\omega_1]} - I_{[0,\omega_2]}) = T(I_{[0,\omega_1]}) - T(I_{[0,\omega_2]}) = Z(\omega_1) - Z(\omega_2).$$

Moreover, since T preserves the norm for any non-intersecting intervals $[\omega_1, \omega_2]$ and $[\omega_3, \omega_4]$ we have

$$\begin{aligned} \operatorname{cov}\left((Z(\omega_1) - Z(\omega_2), (Z(\omega_3) - Z(\omega_4)\right) &= \langle T(I_{[\omega_1,\omega_2]}), T(I_{[\omega_3,\omega_4]})\rangle = \langle I_{[\omega_1,\omega_2]}, I_{[\omega_3,\omega_4]}\rangle \\ &= \int I_{[\omega_1,\omega_2]}(\omega) I_{[\omega_3,\omega_4]}(\omega) dF(\omega) = 0. \end{aligned}$$

Therefore by construction $\{Z(\omega); 0 \leq \omega \leq 2\pi\}$ is an orthogonal increment process, where

$$\begin{aligned} \mathrm{E}|Z(\omega_2) - Z(\omega_1)|^2 &= < T(I_{[\omega_1,\omega_2]}), T(I_{[\omega_1,\omega_2]}) >=< I_{[\omega_1,\omega_2]}, I_{[\omega_1,\omega_2]} > \\ &= \int_0^{2\pi} I_{[\omega_1,\omega_2]} dF(\omega) = \int_{\omega_1}^{\omega_2} dF(\omega) = F(\omega_2) - F(\omega_1). \end{aligned}$$

Having defined the two spaces which are isomorphic and the random function $\{Z(\omega); 0 \leq \omega \leq 2\pi\}$ and function $I_{[0,\omega]}(x)$ which have orthogonal increments, we can now prove the result. Since $dI_{[0,\omega]}(s) = \delta_\omega(s)ds$, where $\delta_\omega(s)$ is the dirac delta function, any function $g \in L_2[0, 2\pi]$ can be

represented as

$$g(\omega) = \int_0^{2\pi} g(s) dI_{[\omega,2\pi]}(s).$$

Thus for $g(\omega) = \exp(-it\omega)$ we have

$$\exp(-it\omega) = \int_0^{2\pi} \exp(-its) dI_{[\omega,2\pi]}(s).$$

Therefore

$$\begin{aligned} T(\exp(-it\omega)) &= T\left(\int_0^{2\pi} \exp(-its) dI_{[\omega,2\pi]}(s)\right) = \int_0^{2\pi} \exp(-its) T[dI_{[\omega,2\pi]}(s)] \\ &= \int_0^{2\pi} \exp(-its) dT[I_{[\omega,2\pi]}(s)], \end{aligned}$$

where the mapping goes inside the integral due to the linearity of the isomorphism. Using that $I_{[\omega,2\pi]}(s) = I_{[0,s]}(\omega)$ we have

$$T(\exp(-it\omega)) = \int_0^{2\pi} \exp(-its) dT[I_{[0,s]}(\omega)].$$

By definition we have $T(I_{[0,s]}(\omega)) = Z(s)$ which we substitute into the above to give

$$X_t = \int_0^{2\pi} \exp(-its) dZ(s),$$

which gives the required result.

Note that there are several different ways to prove this result. □

It is worth taking a step back from the proof and see where the assumption of stationarity crept in. By Bochner's theorem we have that

$$c(t-\tau) = \int \exp(-i(t-\tau)\omega) dF(\omega),$$

where F is a distribution. We use F to define the space $\mathcal{H}_1$, the mapping T (through $\{\exp(ik\omega)\}_k$), the inner-product and thus the isomorphism. However, it was the construction of the orthogonal random functions $\{Z(\omega)\}$ that was instrumental. The main idea of the proof was that there are functions $\{\phi_k(\omega)\}$ and a distribution H such that all the covariances of the stochastic process $\{X_t\}$

can be written as

$$\mathrm{E}(X_t X_\tau) = c(t,\tau) = \int_0^{2\pi} \phi_t(\omega)\overline{\phi_\tau(\omega)} dH(\omega),$$

where H is a measure. As long as the above representation exists, then we can define two spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ where $\{\phi_k\}$ is the basis of the functional space $\mathcal{H}_1$ and it contains all functions f such that $\int |f(\omega)|^2 dH(\omega) < \infty$ and $\mathcal{H}_2$ is the random space defined by $\overline{\mathrm{sp}}(X_t; t \in \mathbb{Z})$. From here we can define an isomorphism $T : \mathcal{H}_1 \to \mathcal{H}_2$, where for all functions $f(\omega) = \sum_k f_k \phi_k(\omega) \in \mathcal{H}_1$

$$T(f) = \sum_k f_k X_k \in \mathcal{H}_2.$$

An important example is $T(\phi_k) = X_k$. Now by using the same arguments as those in the proof above we have

$$X_t = \int \phi_t(\omega) dZ(\omega)$$

where $\{Z(\omega)\}$ are orthogonal random functions and $\mathrm{E}|Z(\omega)|^2 = H(\omega)$. We state this result in the theorem below (see Priestley (1983), Section 4.11).

Theorem 10.9.1 (General orthogonal expansions) *Let $\{X_t\}$ be a time series (not necessarily second order stationary) with covariance $\{\mathrm{E}(X_t X_\tau) = c(t,s)\}$. If there exists a sequence of functions $\{\phi_k(\cdot)\}$ which satisfy for all k*

$$\int_0^{2\pi} |\phi_k(\omega)|^2 dH(\omega) < \infty$$

and the covariance admits the representation

$$c(t,s) = \int_0^{2\pi} \phi_t(\omega)\overline{\phi_s(\omega)} dH(\omega), \tag{10.32}$$

where H is a distribution then for all t we have the representation

$$X_t = \int \phi_t(\omega) dZ(\omega) \tag{10.33}$$

where $\{Z(\omega)\}$ are orthogonal random functions and $\mathrm{E}|Z(\omega)|^2 = H(\omega)$. On the other hand if X_t

has the representation (10.33), then $c(s,t)$ admits the representation (10.32).

Chapter 11

Spectral Analysis

Prerequisites

- The Gaussian likelihood.
- The approximation of a Toeplitz by a Circulant (covered in previous chapters).

Objectives

- The DFTs are close to uncorrelated but have a frequency dependent variance (under stationarity).
- The DFTs are asymptotically Gaussian.
- For a linear time series the DFT is almost equal to the transfer function times the DFT of the innovations.
- The periodograms is the square of the DFT, whose expectation is approximately equal to the spectral density. Smoothing the periodogram leads to an estimator of the spectral density as does truncating the covariances.
- The Whittle likelihood and how it is related to the Gaussian likelihood.
- Understand that many estimator can be written in the frequency domain.
- Calculating the variance of an estimator.

11.1 The DFT and the periodogram

In the previous section we motivated transforming the stationary time series $\{X_t\}$ into it's discrete Fourier transform

$$\begin{aligned} J_n(\omega_k) &= \frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n} X_t \exp(ik\frac{2\pi t}{n}) \\ &= \left(\frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n} X_t \cos(k\frac{2\pi t}{n}) + i\frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n} X_t \sin(k\frac{2\pi t}{n})\right) \qquad k = 0,\ldots,n/2 \end{aligned}$$

(frequency series) as an alternative way of analysing the time series. Since there is a one-to-one mapping between the two, nothing is lost by making this transformation. Our principle reason for using this transformation is given in Lemma 10.2.1, where we showed that $\{J_n(\omega_k)\}_{n=1}^{n/2}$ is an almost uncorrelated series. However, there is a cost to the uncorrelatedness property, that is unlike the original stationary time series $\{X_t\}$, the variance of the DFT varies over the frequencies, and the variance is the spectral density at that frequency. We summarise this result below, but first we recall the definition of the spectral density function

$$f(\omega) = \frac{1}{2\pi}\sum_{r=-\infty}^{\infty} c(r)\exp(ir\omega) \qquad \omega \in [0, 2\pi]. \tag{11.1}$$

We summarize some of the results derived in Chapter 10 here.

Lemma 11.1.1 *Suppose that $\{X_t\}$ is a zero second order stationary time series, where $\mathrm{cov}(X_0, X_r) = c(r)$ and $\sum_r |c(r)| < \infty$. Define $\omega_k = \frac{2\pi k}{n}$. Then*

(i)

$$|J_n(\omega)|^2 = \frac{1}{2\pi}\sum_{r=-(n-1)}^{n-1} \hat{c}_n(r)\exp(ir\omega), \tag{11.2}$$

where $\hat{c}_n(r)$ is the sample autocovariance.

(ii) for $k \neq 0$ we have $\mathrm{E}[J_n(\omega_k)] = 0$,

$$\big|\mathrm{E}(|J_n(\omega)|^2) - f(\omega)\big| \leq \frac{1}{2\pi}\Big(\sum_{|r|\geq n}|c(r)| + \frac{1}{n}\sum_{|r|\leq n}|rc(r)|\Big) \to 0 \tag{11.3}$$

as $n \to \infty$,

(iii)

$$\mathrm{cov}\left[J_n(\frac{2\pi k_1}{n}), J_n(\frac{2\pi k_2}{n})\right] = \begin{cases} f(\frac{2\pi k}{n}) + o(1) & k_1 = k_2 \\ o(1) & k_1 \neq k_2 \end{cases}$$

where $f(\omega)$ is the spectral density function defined in (11.1). Under the stronger condition $\sum_r |rc(r)| < \infty$ the $o(1)$ above is replaced with $O(n^{-1})$.

In addition if we have higher order stationarity (or strict stationarity), then we also can find expressions for the higher order cumulants of the DFT (see Proposition 10.7.1).

It should be noted that even if the mean of the stationary time series $\{X_t\}$ is not zero (ie. $\mathrm{E}(X_t) = \mu \neq 0$), so long as $\omega_k \neq 0$ $\mathrm{E}(J_n(\omega_k)) = 0$ (even without centering X_t, with $X_t - \bar{X}$).

Since there is a one-to-one mapping between the observations and the DFT, it is not surprising that classical estimators can be written in terms of the DFT. For example, the sample covariance can be rewritten in terms of the DFT

$$\widehat{c}_n(r) + \widehat{c}_n(n-r) = \frac{1}{n}\sum_{k=1}^{n} |J_n(\omega_k)|^2 \exp(-ir\omega_k). \tag{11.4}$$

(see Appendix A.3(iv)). Since $\widehat{c}_n(n-r) = \frac{1}{n}\sum_{t=|n-r|}^{n} X_t X_{t+|n-r|}$, for small r (relative to T) this term is negligible, and gives

$$\widehat{c}_n(r) \approx \frac{1}{n}\sum_{k=1}^{n} |J_n(\omega_k)|^2 \exp(-ir\omega_k). \tag{11.5}$$

The modulo square of the DFT plays such an important role in time series analysis that it has it's own name, the periodogram, which is defined as

$$I_n(\omega) = |J_n(\omega)|^2 = \frac{1}{2\pi}\sum_{r=-(n-1)}^{n-1} \hat{c}_n(r)\exp(ir\omega). \tag{11.6}$$

By using Lemma 11.1.1 or Theorem 10.7.1 we have $\mathrm{E}(I_n(\omega)) = f(\omega) + O(\frac{1}{n})$. Moreover, (11.4) belongs to a general class of integrated mean periodogram estimators which have the form

$$A(\phi, I_n) = \frac{1}{n}\sum_{k=1}^{n} I_n(\omega_k)\phi(\omega_k). \tag{11.7}$$

Replacing the sum by an integral and the periodogram by its limit, it is clear that these are

estimators of the integrated spectral density

$$A(f,\phi) = \int_0^{2\pi} f(\omega)\phi(\omega)d\omega.$$

Before we consider these estimators (in Section 11.5). We analyse some of the properties of the DFT.

11.2 Distribution of the DFT and Periodogram under linearity

An interesting aspect of the DFT, is that under certain conditions the DFT is asymptotically normal. We can heuristically justify this by noting that the DFT is a (weighted) sample mean. In fact at frequency zero, it is the sample mean ($J_n(0) = \sqrt{\frac{n}{2\pi}}\bar{X}$). In this section we prove this result, and a similar result for the periodogram. We do the proof under linearity of the time series, that is

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

however the result also holds for nonlinear time series (but is beyond this course).

The DFT of the innovations $J_\varepsilon(\omega_k) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^n \varepsilon_t e^{it\omega_k}$ is a very simple object to deal with it. First the DFT is an orthogonal transformation and the orthogonal transformation of iid random variables leads to uncorrelated random variables. In other words, $\{J_\varepsilon(\omega_k)\}$ is completely uncorrelated as are its real and imaginary parts. Secondly, if $\{\varepsilon_t\}$ are Gaussian, then $\{J_\varepsilon(\omega_k)\}$ are independent and Gaussian. Thus we start by showing the DFT of a linear time series is approximately equal to the DFT of the innovations multiplied by the transfer function. This allows us to transfer results regarding $J_\varepsilon(\omega_k)$ to $J_n(\omega_k)$.

We will use the assumption that $\sum_j |j^{1/2}\psi_j| < \infty$, this is a slightly stronger assumption than $\sum_j |\psi_j| < \infty$ (which we worked under in Chapter 4).

Lemma 11.2.1 *Let us suppose that $\{X_t\}$ satisfy $X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_t$, where $\sum_{j=-\infty}^{\infty} |j^{1/2}\psi_j| < \infty$, and $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2. Let*

$$J_\varepsilon(\omega) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^n \varepsilon_t \exp(it\omega).$$

Then we have

$$J_n(\omega) \;=\; \big\{\sum_j \psi_j \exp(ij\omega)\big\} J_\varepsilon(\omega) + Y_n(\omega), \tag{11.8}$$

where $Y_n(\omega) = \frac{1}{\sqrt{2\pi n}} \sum_j \psi_j \exp(ij\omega) U_{n,j}$, *with* $U_{n,j} = \sum_{t=1-j}^{n-j} \exp(it\omega)\varepsilon_t - \sum_{t=1}^{n} \exp(it\omega)\varepsilon_t$ *and* $\mathrm{E}(Y_n(\omega))^2 \leq (\frac{1}{n^{1/2}} \sum_{j=-\infty}^{\infty} |\psi_j| \min(|j|, n)^{1/2})^2 = O(\frac{1}{n})$.

PROOF. We note that

$$\begin{aligned}
J_n(\omega) &= \sum_{j=-\infty}^{\infty} \psi_j \exp(ij\omega) \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} \varepsilon_{t-j} \exp(it\omega) \\
&= \sum_{j=-\infty}^{\infty} \psi_j \exp(ij\omega) \frac{1}{\sqrt{2\pi n}} \sum_{s=1-j}^{n-j} \varepsilon_s \exp(is\omega) \\
&= \left(\frac{1}{\sqrt{2\pi n}} \sum_j \psi_j \exp(ij\omega) \right) J_\varepsilon(\omega) + \underbrace{\sum_j \psi_j \exp(ij\omega) \left[\sum_{t=1-j}^{n-j} \exp(it\omega)\varepsilon_t - \sum_{t=1}^{n} \exp(it\omega)\varepsilon_t \right]}_{=Y_n(\omega)}.
\end{aligned}$$

We will show that $Y_n(\omega)$ is negligible with respect to the first term. We decompose $Y_n(\omega)$ into three terms

$$\begin{aligned}
Y_n(\omega) &= \frac{1}{\sqrt{2\pi n}} \sum_{j=-\infty}^{-n} \psi_j e^{ij\omega} \underbrace{\left[\sum_{t=1-j}^{n-j} \exp(it\omega)\varepsilon_t - \sum_{t=1}^{n} \exp(it\omega)\varepsilon_t \right]}_{\text{no terms in common}} + \\
&\quad \frac{1}{\sqrt{2\pi n}} \sum_{j=-n}^{n} \psi_j e^{ij\omega} \underbrace{\left[\sum_{t=1-j}^{n-j} \exp(it\omega)\varepsilon_t - \sum_{t=1}^{n} \exp(it\omega)\varepsilon_t \right]}_{(n-j) \text{ terms in common, } 2j \text{ terms not in common}} + \\
&\quad \frac{1}{\sqrt{2\pi n}} \sum_{j=n+1}^{\infty} \psi_j e^{ij\omega} \underbrace{\left[\sum_{t=1-j}^{n-j} \exp(it\omega)\varepsilon_t - \sum_{t=1}^{n} \exp(it\omega)\varepsilon_t \right]}_{\text{no terms in common}} \\
&= I + II + II.
\end{aligned}$$

If we took the expectation of the absolute of $Y_n(\omega)$ we find that we require the condition $\sum_j |j\psi_j| < \infty$ (and we don't exploit independence of the innovations). However, by evaluating $\mathrm{E}|Y_n(\omega)|^2$ we

exploit to independence of $\{\varepsilon_t\}$, ie.

$$\begin{aligned}
[\mathrm{E}(I^2)]^{1/2} &\leq \frac{1}{\sqrt{2\pi n}} \sum_{j=-\infty}^{-n} |\psi_j| \left[\mathrm{E} \left(\sum_{t=1-j}^{n-j} \exp(it\omega)\varepsilon_t - \sum_{t=1}^{n} \exp(it\omega)\varepsilon_t \right)^2 \right]^{1/2} \\
&\leq \frac{1}{\sqrt{2\pi n}} \sum_{j=-\infty}^{-n} |\psi_j| \left[2n\sigma^2\right]^{1/2} \leq \frac{1}{\sqrt{2\pi n}} \sum_{j=-\infty}^{-n} |j^{1/2}\psi_j| \leq \frac{1}{\sqrt{2\pi n}} \sum_{j=-\infty}^{\infty} |j^{1/2}\psi_j|
\end{aligned}$$

similarly, $III = O(n^{-1/2})$ and

$$\begin{aligned}
[\mathrm{E}(I^2)]^{1/2} &\leq \frac{1}{\sqrt{2\pi n}} \sum_{j=-n}^{n} |\psi_j| \left[\mathrm{E} \left(\sum_{t=1-j}^{n-j} \exp(it\omega)\varepsilon_t - \sum_{t=1}^{n} \exp(it\omega)\varepsilon_t \right)^2 \right]^{1/2} \\
&\leq \frac{1}{\sqrt{2\pi n}} \sum_{j=-n}^{-n} |\psi_j| \left[2j\sigma^2\right]^{1/2} \leq \frac{1}{\sqrt{2\pi n}} \sum_{j=-\infty}^{-n} |j^{1/2}\psi_j| \leq \frac{1}{\sqrt{2\pi n}} \sum_{j=-\infty}^{\infty} |j^{1/2}\psi_j|.
\end{aligned}$$

Thus we obtain the desired result. □

The above shows that under linearity and the condition $\sum_j |j^{1/2}\psi_j| < \infty$ we have

$$J_n(\omega) = \Big\{ \sum_j \psi_j \exp(ij\omega) \Big\} J_\varepsilon(\omega) + O_p(\frac{1}{\sqrt{n}}). \tag{11.9}$$

This implies that the distribution of $J_n(\omega)$ is determined by the DFT of the innovations $J_\varepsilon(\omega)$. We generalise the above result to the periodogram.

Lemma 11.2.2 *Let us suppose that $\{X_t\}$ is a linear time series $X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}$, where $\sum_{j=-\infty}^{\infty} |j^{1/2}\psi_j| < \infty$, and $\{\varepsilon_t\}$ are iid random variables with mean zero, variance σ^2 $\mathrm{E}(\varepsilon_t^4) < \infty$. Then we have*

$$I_n(\omega) = \left| \sum_j \psi_j \exp(ij\omega) \right|^2 |J_\varepsilon(\omega)|^2 + R_n(\omega), \tag{11.10}$$

where $\mathrm{E}(\sup_\omega |R_n(\omega)|) = O(\frac{1}{n})$.

PROOF. See Priestley (1983), Theorem 6.2.1 or Brockwell and Davis (1998), Theorem 10.3.1. □

To summarise the above result, for a general linear process $X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}$ we have

$$I_n(\omega) = |\sum_j \psi_j \exp(ij\omega)|^2 |J_\varepsilon(\omega)|^2 + O_p(\frac{1}{n}) = 2\pi f(\omega) I_\varepsilon(\omega) + O_p(\frac{1}{n}), \tag{11.11}$$

where we assume w.l.o.g. that $\mathrm{var}(\varepsilon_t) = 1$ and $f(\omega) = \frac{1}{2\pi}|\sum_j \psi_j \exp(ij\omega)|^2$ is the spectral density of $\{X_t\}$.

The asymptotic normality of $J_n(\omega)$ follows from asymptotic normality of $J_\varepsilon(\omega)$, which we prove in the following proposition.

Proposition 11.2.1 *Suppse $\{\varepsilon_t\}$ are iid random variables with mean zero and variance σ^2. We define $J_\varepsilon(\omega) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^n \varepsilon_t \exp(it\omega)$ and $I_\varepsilon(\omega) = \frac{1}{2\pi n}\left|\sum_{t=1}^n \varepsilon_t \exp(it\omega)\right|^2$. Then we have*

$$\underline{J}_\varepsilon(\omega) = \begin{pmatrix} \Re J_\varepsilon(\omega) \\ \Im J_\varepsilon(\omega) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\sigma^2}{2(2\pi)} I_2\right), \tag{11.12}$$

where I_2 is the identity matrix. Furthermore, for any finite m

$$(\underline{J}_\varepsilon(\omega_{k_1})', \ldots, \underline{J}_\varepsilon(\omega_{k_m})') \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\sigma^2}{2(2\pi)} I_{2m}\right), \tag{11.13}$$

$I_\varepsilon(\omega)/\sigma^2 \xrightarrow{\mathcal{D}} \chi^2(2)/2$ (which is equivalent to the exponential distribution with mean one) and

$$\mathrm{cov}(|J_\varepsilon(\omega_j)|^2, |J_\varepsilon(\omega_k)|^2) = \begin{cases} \frac{\kappa_4}{(2\pi)^2 n} & j \neq k \\ \frac{\kappa_4}{(2\pi)^2 n} + \frac{2\sigma^4}{(2\pi)^2} & j = k \end{cases} \tag{11.14}$$

where $\omega_j = 2\pi j/n$ and $\omega_k = 2\pi k/n$ (and $j, k \neq 0$ or n).

PROOF. We first show (11.15). We note that $\Re(J_\varepsilon(\omega_k)) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^n \alpha_{t,n}$ and $\Im(J_\varepsilon(\omega_k)) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^n \beta_{t,n}$ where $\alpha_{t,n} = \varepsilon_t \cos(2k\pi t/n)$ and $\beta_{t,n} = \varepsilon_t \sin(2k\pi t/n)$. We note that $\Re(J_\varepsilon(\omega_k)) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^n \alpha_{t,n}$ and $\Im(J_\varepsilon(\omega_k)) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^n \beta_{t,n}$ are the weighted sum of iid random variables, hence $\{\alpha_{t,n}\}$ and $\{\beta_{t,n}\}$ are martingale differences. Therefore, to show asymptotic normality, we will use the martingale central limit theorem with the Cramer-Wold device to show that (11.15). We note that since $\{\alpha_{t,n}\}$ and $\{\beta_{t,n}\}$ are independent random variables we an prove the same result using a CLT for independent, non-identically distributed variables. However, for practice we will use a martingale CLT. To prove the result we need to verify the three conditions of the martingale

CLT. First we consider the conditional variances

$$\begin{aligned}
\frac{1}{2\pi n}\sum_{t=1}^{n}\mathrm{E}\big(|\alpha_{t,n}|^2\big|\varepsilon_{t-1},\varepsilon_{t-2},\ldots,\varepsilon_1\big) &= \frac{1}{2\pi n}\sum_{t=1}^{n}\cos(2k\pi t/n)^2\varepsilon_t^2 \xrightarrow{\mathcal{P}} \frac{\sigma^2}{2\pi}\\
\frac{1}{2\pi n}\sum_{t=1}^{n}\mathrm{E}\big(|\beta_{t,n}|^2\big|\varepsilon_{t-1},\varepsilon_{t-2},\ldots,\varepsilon_1\big) &= \frac{1}{2\pi n}\sum_{t=1}^{n}\sin(2k\pi t/n)^2\varepsilon_t^2 \xrightarrow{\mathcal{P}} \frac{\sigma^2}{2\pi}\\
\frac{1}{2\pi n}\sum_{t=1}^{n}\mathrm{E}\big(\alpha_{t,n}\beta_{t,n}\big|\varepsilon_{t-1},\varepsilon_{t-2},\ldots,\varepsilon_1\big) &= \frac{1}{2\pi n}\sum_{t=1}^{n}\cos(2k\pi t/n)\sin(2k\pi t/n)\varepsilon_t^2 \xrightarrow{\mathcal{P}} 0,
\end{aligned}$$

where the above follows from basic calculations using the mean and variance of the above. Finally we need to verify the Lindeberg condition, we only verify it for $\frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n}\alpha_{t,n}$, the same argument holds true for $\frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n}\beta_{t,n}$. We note that for every $\epsilon>0$ we have

$$\frac{1}{2\pi n}\sum_{t=1}^{n}\mathrm{E}\big(|\alpha_{t,n}|^2 I(|\alpha_{t,n}|\geq 2\pi\sqrt{n}\epsilon)\big|\varepsilon_{t-1},\varepsilon_{t-2},\ldots\big)=\frac{1}{2\pi n}\sum_{t=1}^{n}\mathrm{E}\left[|\alpha_{t,n}|^2 I(|\alpha_{t,n}|\geq 2\pi\sqrt{n}\epsilon)\right].$$

By using $|\alpha_{t,n}|=|\cos(2\pi t/n)\varepsilon_t|\leq|\varepsilon_t|$ the above can be bounded by

$$\begin{aligned}
&\frac{1}{2\pi n}\sum_{t=1}^{n}\mathrm{E}\big[|\alpha_{t,n}|^2 I(|\alpha_{t,n}|\geq 2\pi\sqrt{n}\epsilon)\big]\\
\leq\ &\frac{1}{2\pi n}\sum_{t=1}^{n}\mathrm{E}\big[|\varepsilon_t|^2 I(|\varepsilon_t|\geq 2\pi\sqrt{n}\epsilon)\big]=\mathrm{E}\big[|\varepsilon_t|^2 I(|\varepsilon_t|\geq 2\pi\sqrt{n}\epsilon)\big]\xrightarrow{\mathcal{P}}0 \quad \text{as} \quad n\to\infty,
\end{aligned}$$

the above is true because $\mathrm{E}(\varepsilon_t^2)<\infty$. Hence we have verified Lindeberg condition and we obtain (11.15). The proof of (11.13) is similar, hence we omit the details. Because $I_\varepsilon(\omega)=\Re(J_\varepsilon(\omega))^2+\Im(J_\varepsilon(\omega))^2$, from (11.15) we have $I_\varepsilon(\omega)/\sigma^2\sim\chi^2(2)/2$ (which is the same as an exponential with mean one).

To prove (11.14) we can either derive it from first principles or by using Proposition 10.7.1. Here we do it from first principles. We observe

$$\mathrm{cov}(I_\varepsilon(\omega_j),I_\varepsilon(\omega_k)) = \frac{1}{(2\pi)^2 n^2}\sum_{k_1}\sum_{k_2}\sum_{t_1}\sum_{t_2}\mathrm{cov}(\varepsilon_{t_1}\varepsilon_{t_1+k_1},\varepsilon_{t_2}\varepsilon_{t_2+k_2}).$$

Expanding the covariance gives

$$\begin{aligned}
\mathrm{cov}(\varepsilon_{t_1}\varepsilon_{t_1+k_1},\varepsilon_{t_2}\varepsilon_{t_2+k_2}) = \ &\mathrm{cov}(\varepsilon_{t_1},\varepsilon_{t_2+k_2})\mathrm{cov}(\varepsilon_{t_2},\varepsilon_{t_1+k_1})+\mathrm{cov}(\varepsilon_{t_1},\varepsilon_{t_2})\mathrm{cov}(\varepsilon_{t_1+k_1},\varepsilon_{t_2+k_2})+\\
&\mathrm{cum}(\varepsilon_{t_1},\varepsilon_{t_1+k_1},\varepsilon_{t_2},\varepsilon_{t_2+k_2}).
\end{aligned}$$

Since $\{\varepsilon_t\}$ are iid random variables, for most t_1, t_2, k_1 and k_2 the above covariance is zero. The exceptions are when $t_1 = t_2$ and $k_1 = k_2$ or $t_1 = t_2$ and $k_1 = k_2 = 0$ or $t_1 - t_2 = k_1 = -k_2$. Counting all these combinations we have

$$\mathrm{cov}(|J_\varepsilon(\omega_j)|^2, |J_\varepsilon(\omega_k)|^2) = \frac{2\sigma^4}{(2\pi)^2 n^2} \sum_k \sum_t \sum_t \exp(ik(\omega_j - \omega_k)) + \frac{1}{(2\pi)^2 n^2} \sum_t \kappa_4$$

where $\sigma^2 = \mathrm{var}(\varepsilon_t)$ and $\kappa_4 = \mathrm{cum}_4(\varepsilon) = \mathrm{cum}(\varepsilon_t, \varepsilon_t, \varepsilon_t, \varepsilon_t)$. We note that for $j \neq k$, $\sum_t \exp(ik(\omega_j - \omega_k)) = 0$ and for $j = k$, $\sum_t \exp(ik(\omega_j - \omega_k)) = n$, substutiting this into $\mathrm{cov}(|J_\varepsilon(\omega_j)|^2, |J_\varepsilon(\omega_k)|^2)$ gives us the desired result. □

By using (11.9) the following result follows immediately from Lemma 11.2.1, equation (11.15).

Corollary 11.2.1 *Let us suppose that $\{X_t\}$ is a linear time series $X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}$, where $\sum_{j=-\infty}^{\infty} |j^{1/2}\psi_j| < \infty$, and $\{\varepsilon_t\}$ are iid random variables with mean zero, variance σ^2 $\mathrm{E}(\varepsilon_t^4) < \infty$. Then we have*

$$\begin{pmatrix} \Re J_n(\omega) \\ \Im J_n(\omega) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{2} f(\omega) I_2\right), \tag{11.15}$$

Using (11.11) we see that $I_n(\omega) \approx f(\omega)|J_\varepsilon(\omega)|^2$. This suggest that most of the properties which apply to $|J_\varepsilon(\omega)^2$ also apply to $I_n(\omega)$. Indeed in the following theorem we show that the asympototic distribution of $I_n(\omega)$ is exponential with asymptotic mean $f(\omega)$ and variance $f(\omega)^2$ (unless $\omega = 0$ in which case it is $2f(\omega)^2$).

By using Lemma 11.2.1 we now generalise Proposition 11.2.1 to linear processes. We show that just like the DFT the Periodogram is also 'near uncorrelated' at different frequencies. This result will be useful when motivating and deriving the sampling of the spectral density estimator in Section 11.3.

Theorem 11.2.1 *Suppose $\{X_t\}$ is a linear time series $X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}$, where $\sum_{j=-\infty}^{\infty} |j^{1/2}\psi_j| < \infty$ with $\mathrm{E}[\varepsilon_t] = 0$, $\mathrm{var}[\varepsilon_t] = \sigma^2$ and $\mathrm{E}[\varepsilon_t^4] < \infty$. Let $I_n(\omega)$ denote the periodogram associated with $\{X_1, \ldots, X_n\}$ and $f(\cdot)$ be the spectral density. Then*

(i) If $f(\omega) > 0$ for all $\omega \in [0, 2\pi]$ and $0 < \omega_1, \ldots, \omega_m < \pi$, then

$$\big(I_n(\omega_1)/f(\omega_1), \ldots, I_n(\omega_m)/f(\omega_m)\big)$$

converges in distribution (as $n \to \infty$) to a vector of independent exponential distributions with mean one.

(ii) Furthermore, for $\omega_j = \frac{2\pi j}{n}$ and $\omega_k = \frac{2\pi k}{n}$ we have

$$\mathrm{cov}(I_n(\omega_k), I_n(\omega_j)) = \begin{cases} 2f(\omega_k)^2 + O(n^{-1/2}) & \omega_j = \omega_k = 0 \text{ or } \pi \\ f(\omega_k)^2 + O(n^{-1/2}) & 0 < \omega_j = \omega_k < \pi \\ O(n^{-1}) & \omega_j \neq \omega_k \end{cases}$$

where the bound is uniform in ω_j and ω_k.

Remark 11.2.1 (Summary of properties of the periodogram) *(i) The periodogram is non-negative and is an asymptotically an unbiased estimator of the spectral density (when $\sum_j |\psi_j| < \infty$).*

(ii) It symmetric about zero, $I_n(\omega) = I_n(\omega + \pi)$, like the spectral density function.

(iii) At the fundemental frequencies $\{I_n(\omega_j)\}$ are asymptotically uncorrelated.

(iv) If $0 < \omega < \pi$, $I_n(\omega)$ is asymptotically exponentially distributed with mean $f(\omega)$.

It should be mentioned that Theorem 11.2.1 also holds for several nonlinear time series too.

11.3 Estimating the spectral density function

There are several explanations as to why the raw periodogram can not be used as an estimator of the spectral density function, despite its mean being approximately equal to the spectral density. One explanation is a direct consequence of Theorem 11.2.1, where we showed that the distribution of the periodogram standardized with the spectral density function is an exponential distribution, from here it is clear it will not converge to the mean, however large the sample size. An alternative, explanation is that the periodogram is the Fourier transform of the autocovariances estimators at n different lags. Typically the variance for each covariance $\hat{c}_n(k)$ will be about $O(n^{-1})$, thus, roughly speaking, the variance of $I_n(\omega)$ will be the sum of these n $O(n^{-1})$ variances which leads to a variance of $O(1)$, this clearly does not converge to zero.

Both these explanation motivate estimators of the spectral density function, which turn out to be the same. It is worth noting that Parzen (1957) first proposed a consistent estimator of the

spectral density. These results not only lead to a revolution in spectral density estimation but also the usual density estimation that you may have encountered in nonparametric statistics (one of the first papers on density estimation is Parzen (1962)).

We recall that $J_n(\omega_k)$ are zero mean uncorrelated random variables whose variance is almost equal to $f(\omega_k)$. This means that $\mathrm{E}|J_n(\omega_k)|^2 = \mathrm{E}[I_n(\omega_n)] \approx f(\omega_k)$.

Remark 11.3.1 (Smoothness of the spectral density) *We observe that*

$$f^{(s)}(\omega) = \frac{1}{(2\pi)} \sum_{r\in\mathbb{Z}} (ir)^s c(r) \exp(ir\omega).$$

Therefore, the smoothness of the spectral density function is determined by finiteness of $\sum_r |r^s c(r)|$, *in other words how fast the autocovariance function converges to zero. We recall that the acf of ARMA processes decay exponential fast to zero, thus* f *is extremely smooth (all derivatives exist).*

Assuming that the autocovariance function converges to zero sufficiently fast f will slowly vary over frequency. Furthermore, using Theorem 11.2.1, we know that $\{I_n(\omega_k)\}$ are close to uncorrelated and $I_n(\omega_k)/f(\omega_k)$ is $2^{-1}\chi^2_2$. Therefore we can write $I_n(\omega_k)$ as

$$\begin{aligned} I_n(\omega_k) &= \mathrm{E}(I_n(\omega_k)) + [I_n(\omega_k) - \mathrm{E}(I_n(\omega_k))] \\ &\approx f(\omega_k) + f(\omega_k)U_k, \qquad k = 1, \ldots, n, \end{aligned} \tag{11.16}$$

where $\{U_k\}$ is sequence of mean zero and constant variance almost uncorrelated random variables.

We recall (11.16) resembles the usual nonparametric equation (function plus noise) often considered in nonparametric statistics.

Remark 11.3.2 (Nonparametric Kernel estimation) *Let us suppose that we observe* Y_i *where*

$$Y_i = g\left(\frac{i}{n}\right) + \varepsilon_i \qquad 1 \leq i \leq n,$$

and $\{\varepsilon_i\}$ *are iid random variables and* $g(\cdot)$ *is a 'smooth' function. The kernel density estimator of* $\widehat{g}_n(\frac{i}{n})$

$$\widehat{g}_n\left(\frac{j}{n}\right) = \sum_i \frac{1}{bn} W\left(\frac{j-i}{bn}\right) Y_i,$$

where $W(\cdot)$ *is a smooth kernel function of your choosing, such as the Gaussian kernel, etc.*

This suggest that to estimate the spectral density we could use a local weighted average of $\{I_n(\omega_k)\}$. Equation (11.16) motivates the following nonparametric estimator of $f(\omega)$

$$\widehat{f}_n(\omega_j) = \sum_k \frac{1}{bn} W\left(\frac{j-k}{bn}\right) I_n(\omega_k), \tag{11.17}$$

where $W(\cdot)$ is a spectral window which satisfies $\int W(x)dx = 1$ and $\int W(x)^2 dx < \infty$.

Example 11.3.1 (Spectral windows) *Here we give examples of spectral windows (see Section 6.2.3, page 437 in Priestley (1983)).*

(i) The Daniell spectral Window is the local average

$$W(x) = \begin{cases} 1/2 & |x| \leq 1 \\ 0 & |x| > 1 \end{cases}$$

This window leads to the estimator

$$\widehat{f}_n(\omega_j) = \frac{1}{bn} \sum_{k=j-bn/2}^{j+bn/2} I_n(\omega_k).$$

A plot of the periodgram, spectral density and different estimators (using Daniell kernel with $bn = 2$ and $bn = 10$) of the AR(2) process $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$ is given in Figure 11.1. We observe that too small b leads to undersmoothing but too large b leads to over smoothing of features. There are various methods for selecting the bandwidth, one commonly method based on the Kullbach-Leibler criterion is proposed in Beltrao and Bloomfield (1987).

(ii) The Bartlett-Priestley spectral Window

$$W(x) = \begin{cases} \frac{3}{4}\left(1 - x^2\right) & |x| \leq 1 \\ 0 & |x| > 1 \end{cases}$$

This spectral window was designed to reduce the mean squared error of the spectral density estimator (under certain smoothness conditions).

The above estimator was constructed within the frequency domain. We now consider a spectral density estimator constructed within the 'time domain'. We do this by considering the periodogram

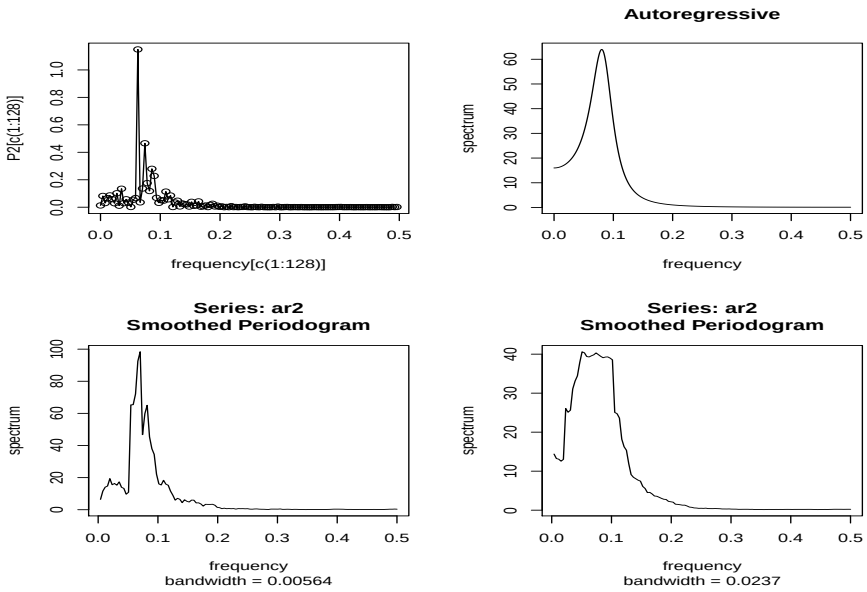


Figure 11.1: Using a realisation of the AR(2): $X_t = 1.5X_{t-1} - 0.75X_{t-2} + \varepsilon_t$ where $n = 256$. Top left: Periodogram, Top Right: True spectral density function. Bottom left: Spectral density estimator with $bn = 2$ and Bottom right: Spectral density estimator with $bn = 10$.

from an alternative angle. We recall that

$$I_n(\omega) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \hat{c}_n(k) \exp(ik\omega),$$

thus it is the sum of n autocovariance estimators. This is a type of sieve estimator (a nonparametric function estimator which estimates the coefficients/covariances in a series expansion). But as we explained above, this estimator is not viable because it uses too many coefficient estimators. Since the true coefficients/covariances decay to zero for large lags, this suggests that we do not use all the sample covariances in the estimator, just some of them. Hence a viable estimator of the spectral density is the truncated autocovariance estimator

$$\widetilde{f}_n(\omega) = \frac{1}{2\pi} \sum_{k=-m}^{m} \hat{c}_n(k) \exp(ik\omega), \tag{11.18}$$

or a generalised version of this which down weights the sample autocovariances at larger lags

$$\widetilde{f}_n(\omega) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \lambda\left(\frac{k}{m}\right) \hat{c}_n(k) \exp(ik\omega), \tag{11.19}$$

where $\lambda(\cdot)$ is the so called lag window. The estimators (11.17) and (11.19) are very conceptionally similar, this can be understood if we rewrite $\hat{c}_n(r)$ in terms the periodogram $\hat{c}_n(r) = \int_0^{2\pi} I_n(\omega)\exp(-ir\omega)d\omega$, and transforming (11.19) back into the frequency domain

$$\widetilde{f}_n(\omega) = \frac{1}{2\pi}\int I_n(\lambda)\sum_{k=-(n-1)}^{n-1}\lambda(\frac{k}{m})\exp(ik(\omega-\lambda))d\lambda = \frac{1}{2\pi}\int I_n(\lambda)W_m(\omega-\lambda)d\omega, \quad (11.20)$$

where $W_m(\omega) = \frac{1}{2\pi}\sum_{k=-(n-1)}^{n-1}\lambda(\frac{k}{m})\exp(ik\omega)$.

Example 11.3.2 (Examples of Lag windows) *Here we detail examples of lag windows.*

(i) *Truncated Periodogram lag Window* $\lambda(u) = I_{[-1,1]}(u)$, *where* $\{\lambda(k/m)\}$ *corresponds to*

$$W_m(x) = \frac{1}{2\pi}\sum_{k=-m}^{m} e^{ik\omega} = \frac{1}{2\pi}\frac{\sin[(m+1/2)x]}{\sin(x/2)},$$

which is the Dirichlet kernel.

Note that the Dirchlet kernel can be negative, thus we can see from (11.20) that $\widetilde{f}_n$ *can be negative. Which is one potential drawback of this estimator (see Example 10.6.2).*

(ii) *The Bartlett lag Window* $\lambda(x) = (1-|x|)I_{[-1,1]}(x)$, *where* $\{\lambda(k/m)\}$ *corresponds to*

$$W_m(x) = \frac{1}{2\pi}\sum_{k=-m}^{m}\left(1-\frac{||k}{m}\right)e^{ik\omega} = \frac{1}{2\pi n}\left(\frac{\sin(nx/2)}{\sin(x/2)}\right)^2$$

which is the Fejer kernel. We can immediately see that one advantage of the Bartlett window is that it corresponds to a spectral density estimator which is positive.

Note that in the case that $m = n$ *(the sample size), the truncated periodogram window estimator corresponds to* $\sum_{|r|\leq n} c(r)e^{ir\omega}$ *and the Bartlett window estimator corresponds to* $\sum_{|r|\leq n}[1-|r|/n]c(r)e^{ir\omega}$.

$W_m(\cdot)$ and $\frac{1}{b}W(\frac{\cdot}{b})$ (defined in (11.17)) cannot not be the same function, but they share many of the same characteristics. In particular,

$$\begin{aligned} W_m(\omega) &= \sum_{k=-(n-1)}^{n-1}\lambda\left(\frac{k}{m}\right)\exp(ik\omega) = m\sum_{k=-(m-1)}^{m-1}\frac{1}{m}\lambda\left(\frac{k}{m}\right)\exp\left(i\frac{k}{m}\cdot m\omega\right) \\ &= m\frac{1}{m}\sum_{k=-(m-1)}^{m-1}\lambda\left(\omega_k\right)\exp\left(i\omega_k(m\omega)\right), \end{aligned}$$

where $\omega_k = k/n$. By using (A.2) and (A.3) (in the appendix), we can approximate the sum by the integral and obtain

$$W_m(\omega) \quad = \quad mW(m\omega) + O(1), \qquad \text{where } W(\omega) = \int \lambda(x)\exp(i\omega)dx.$$

Therefore

$$\widetilde{f}_n(\omega) \quad \approx \quad m\int I_n(\lambda)K(m(\omega-\lambda))d\omega.$$

Comparing with $\widehat{f}_n$ and $\widetilde{f}_n(\omega)$ we see that m plays the same role as b^{-1}. Furthermore, we observe $\sum_k \frac{1}{bn}W(\frac{j-k}{bn})I(\omega_k)$ is the sum of about nb $I(\omega_k)$ terms. The equivalent for $W_m(\cdot)$, is that it has the 'spectral' width n/m. In other words since $\tilde{f}_n(\omega) = \frac{1}{2\pi}\sum_{k=-(n-1)}^{n-1}\lambda(\frac{k}{M})\hat{c}_n(k)\exp(ik\omega) = \frac{1}{2\pi}\int mI_n(\lambda)W(M(\omega-\lambda))d\omega$, it is the sum of about n/m terms.

We now analyze the sampling properties of the spectral density estimator. It is worth noting that the analysis is very similar to the analysis of nonparametric kernel regression estimator $\widehat{g}_n(\frac{j}{n}) = \frac{1}{bn}\sum_i W(\frac{j-i}{bn})Y_i$, where $Y_i = g(\frac{i}{n}) + g(\frac{i}{n})\varepsilon_i$ and $\{\varepsilon_i\}$ are iid random variables. This is because the periodogram $\{I_n(\omega)\}_k$ is 'near uncorrelated'. However, still some care needs to be taken in the proof to ensure that the errors in the near uncorrelated term does not build up.

Theorem 11.3.1 *Suppose $\{X_t\}$ satisfy $X_t = \sum_{j=-\infty}^{\infty}\psi_j\varepsilon_{t-j}$, where $\sum_{j=-\infty}^{\infty}|j\psi_j| < \infty$ and $\mathrm{E}(\varepsilon_t^4) < \infty$. Let $\hat{f}_n(\omega)$ be the spectral estimator defined in (11.17). Then*

$$\left|\mathrm{E}(\hat{f}_n(\omega_j)) - f(\omega_j)\right| \leq C\left(\frac{1}{n} + b\right) \tag{11.21}$$

and

$$\mathrm{var}[\hat{f}_n(\omega_j)] \to \begin{cases} \frac{1}{bn}f(\omega_j)^2 & 0 < \omega_j < \pi \\ \frac{2}{bn}f(\omega_j)^2 & \omega_j = 0 \text{ or } \pi \end{cases}, \tag{11.22}$$

$bn \to \infty$, $b \to 0$ as $n \to \infty$.

PROOF. The proof of both (11.21) and (11.22) are based on the spectral window $W(x/b)$ becoming narrower as $b \to 0$, hence there is increasing localisation as the sample size grows (just like nonparametric regression).

We first note that by using Lemma 6.1.1(ii) we have $\sum_r |rc(r)| < \infty$, thus $|f'(\omega)| \leq \sum_r |rc(r)| <$

∞. Hence f is continuous with a bounded first derivative.

To prove (11.21) we take expections

$$\begin{aligned}\Big|\mathrm{E}(\hat{f}_n(\omega_j)) - f(\omega_j)\Big| &= \left|\sum_k \frac{1}{bn} W\left(\frac{k}{bn}\right)\big\{\mathrm{E}\big[I(\omega_{j-k})\big] - f(\omega_j)\big\}\right| \\ &= \sum_k \frac{1}{bn}\left|W\left(\frac{k}{bn}\right)\right| \big|\mathrm{E}\big[I(\omega_{j-k})\big] - f(\omega_{j-k})\big| + \sum_k \frac{1}{bn}\left|W\left(\frac{k}{bn}\right)\right| |f(\omega_j) - f(\omega_{j-k})| \\ &:= I + II.\end{aligned}$$

Using Lemma 11.1.1 we have

$$\begin{aligned}I &= \sum_k \frac{1}{bn}\left|W\left(\frac{k}{bn}\right)|\right| \big|\mathrm{E}\big(I(\omega_{j-k})\big) - f(\omega_{j-k})\big| \\ &\leq C\left(\frac{1}{bn}\sum_k |W(\frac{k}{bn})|\right)\left(\sum_{|k|\geq n}|c(k)| + \frac{1}{n}\sum_{|k|\leq n}|kc(k)|\right) = O(\frac{1}{n}).\end{aligned}$$

To bound II we use that $|f(\omega_1) - f(\omega_2)| \leq \sup|f'(\omega)| \cdot |\omega_1 - \omega_2|$, this gives

$$II = \big|\sum_k \frac{1}{bn} K(\frac{k}{bn})\big\{f(\omega_j) - f(\omega_{j-k})\big\}\big| = O(b).$$

Altogether this gives $I = O(n^{-1})$ and $II = O(b)$ as $bn \to \infty$, $b \to 0$ and $n \to \infty$. The above two bounds mean give (11.21).

We will use Theorem 11.2.1 to prove (11.22). We first assume that $j \neq 0$ or n. To prove the result we use that

$$\begin{aligned}&\mathrm{cov}(|J_n(\omega_{k_1})|^2, |J_n(\omega_{k_2})|^2) = \\ &[f(\omega_{k_1})I(k_1 = k_2) + O(\frac{1}{n})]^2 + [f(\omega_{k_1})I(k_1 = n - k_2) + O(\frac{1}{n})][f(\omega_{k_1})I(n - k_1 = k_2) + O(\frac{1}{n})] \\ &+[\frac{1}{n}f_4(\omega_1, -\omega_1, \omega_2) + O(\frac{1}{n^2})].\end{aligned}$$

where the above follows from Proposition 10.7.1. This gives

$$
\begin{aligned}
& \operatorname{var}(\widehat{f}_n(\omega_j)) \\
= & \sum_{k_1,k_2} \frac{1}{(bn)^2} W\left(\frac{j-k_1}{bn}\right) W\left(\frac{j-k_2}{bn}\right) \operatorname{cov}(I(\omega_{k_1}), I(\omega_{k_1})) \\
= & \sum_{k_1,k_2} \frac{1}{(bn)^2} W\left(\frac{j-k_1}{bn}\right) W\left(\frac{j-k_2}{bn}\right) \\
& \Big(\big[f(\omega_{k_1}) I(k_1 = k_2) + O(\frac{1}{n})\big]^2 + \big[f(\omega_{k_1}) I(k_1 = n - k_2) + O(\frac{1}{n})\big][f(\omega_{k_1}) I(n - k_1 = k_2) + O(\frac{1}{n})\big] \\
& + \big[\frac{1}{n} f_4(\omega_1, -\omega_1, \omega_2) + O(\frac{1}{n^2})\big] \Big) \\
= & \sum_{k=1}^{n} \frac{1}{(bn)^2} W\left(\frac{j-k_1}{bn}\right)^2 f(\omega_k^2) \\
& + \sum_{k=1}^{n} \frac{1}{(bn)^2} W\left(\frac{j-k_1}{bn}\right) W\left(\frac{j-(n-k_1)}{bn}\right) f(\omega_k^2) + O(\frac{1}{n}) \\
= & \frac{1}{2\pi nb} \int \frac{1}{b} W\left(\frac{\omega_j - \omega}{b}\right)^2 f(\omega)^2 d\omega + \frac{1}{2\pi nb} \underbrace{\int \frac{1}{b} W\left(\frac{\omega_j - 2\pi + \omega}{b}\right) W\left(\frac{\omega_j - \omega}{b}\right) f(\omega) d\omega}_{\to 0} + O(\frac{1}{n}) \\
= & \frac{1}{2\pi nb} f(\omega_j)^2 \int \frac{1}{b} W\left(\frac{\omega}{b}\right)^2 d\omega + O(\frac{1}{n})
\end{aligned}
$$

where the above is using the Riemann integral. A similar proof can be used to prove the case $j = 0$ or n. □

The above result means that the mean squared error of the estimator

$$\mathrm{E}\big[\hat{f}_n(\omega_j) - f(\omega_j)\big]^2 \to 0,$$

where $bn \to \infty$ and $b \to 0$ as $n \to \infty$. Moreover

$$\mathrm{E}\big[\hat{f}_n(\omega_j) - f(\omega_j)\big]^2 = O\left(\frac{1}{bn} + b^2\right).$$

Remark 11.3.3 (The distribution of the spectral density estimator) *Using that the periodogram $I_n(\omega)/f(\omega)$ is asymptotically exponentially distributed and uncorrelated at the fundemental frequencies, we can heuristically deduce the limiting distribution of $\hat{f}_n(\omega)$. Here we consider the*

distribution with the rectangular spectral window

$$\hat{f}_n(\omega_j) = \frac{1}{bn}\sum_{k=j-bn/2}^{j+bn/2} I(\omega_k).$$

Since $I(\omega_k)/f(\omega_k)$ are approximately $\chi^2(2)/2$, then since the sum $\sum_{k=j-bn/2}^{j+bn/2} I(\omega_k)$ is taken over a local neighbourhood of ω_j, we have that $f(\omega_j)^{-1}\sum_{k=j-bn/2}^{j+bn/2} I(\omega_k)$ is approximately $\chi^2(2bn)/2$.

We note that when bn is large, then $\chi^2(2bn)/2$ is close to normal. Hence

$$\sqrt{bn}\hat{f}_n(\omega_j) \approx N(f(\omega_j), f(\omega_j)^2).$$

Using this these asymptotic results, we can construct confidence intervals for $f(\omega_j)$.

In general, to prove normality of $\hat{f}_n$ we rewrite it as a quadratic form, from this asymptotic normality can be derived, where

$$\sqrt{bn}\hat{f}_n(\omega_j) \approx N\bigg(f(\omega_j), f(\omega_j)^2\int W(u)^2du\bigg).$$

The variance of the spectral density estimator is simple to derive by using Proposition 10.7.1. The remarkable aspect is that the variance of the spectral density does not involve (asymptotically) the fourth order cumulant (as it is off lower order).

11.4 The Whittle Likelihood

In Chapter 8 we considered various methods for estimating the parameters of an ARMA process. The most efficient method (in terms of Fisher efficiency), when the errors are Gaussian is the Gaussian maximum likelihood estimator. This estimator was defined in the time domain, but it is interesting to note that a very similar estimator which is asymptotically equivalent to the GMLE estimator can be defined within the frequency domain. We start by using heuristics to define the Whittle likelihood. We then show how it is related to the Gaussian maximum likelihood.

To motivate the method let us return to the Sunspot data considered in Exercise 5.1. The

Periodogram and the spectral density corresponding to the best fitting autoregressive model,

$$f(\omega) = (2\pi)^{-1}\Big|1 - 1.1584e^{i\omega} - 0.3890e^{i2\omega} - 0.1674e^{i3\omega} - 0.1385e^{i4\omega} - 0.1054e^{i5\omega} - 0.0559e^{i6\omega} - 0.0049e^{i7\omega} - 0.0572e^{i8\omega} - 0.2378e^{\omega}\Big|^{-2},$$

is given in Figure 11.2. We see that the spectral density of the best fitting AR process closely follows the shape of the periodogram (the DFT modulo square). This means that indirectly the autoregressive estimator (Yule-Walker) chose the AR parameters which best fitted the shape of the periodogram. The Whittle likelihood estimator, that we describe below, does this directly. By selecting the parametric spectral density function which best fits the periodogram. The Whittle

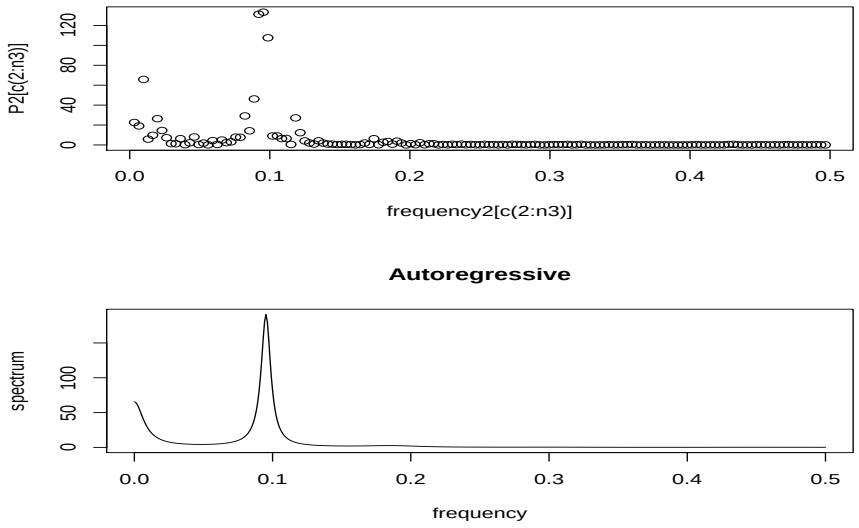


Figure 11.2: The periodogram of sunspot data (with the mean removed, which is necessary to prevent a huge peak at zero) and the spectral density of the best fitting AR model.

likelihood measures the distance between $I_n(\omega)$ and the parametric spectral density function using the Kullbach-Leibler criterion

$$L_n^w(\boldsymbol{\theta}) = \sum_{k=1}^{n}\left(\log f_{\boldsymbol{\theta}}(\omega_k) + \frac{I_n(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)}\right), \quad \omega_k = \frac{2\pi k}{n},$$

and the parametric model which minimises this 'distance' is used as the estimated model. The choice of this criterion over the other distance criterions may appear to be a little arbitrary, however there are several reasons why this is considered a good choice. Below we give some justifications as to

why this criterion is the prefered one.

First let us suppose that we observe $\{X_t\}_{t=1}^n$, where X_t satisfies the ARMA representation

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

and $\{\varepsilon_t\}$ are iid random variables. We will assume that $\{\phi_j\}$ and $\{\psi_j\}$ are such that the roots of their corresponding characteristic polynomial are greater than $1+\delta$. Let $\boldsymbol{\theta} = (\underline{\phi}, \underline{\theta})$. As we mentioned in Section 10.2 if $\sum_r |rc(r)| < \infty$, then

$$\operatorname{cov}(J_n(\omega_{k_1}), J_n(\omega_{k_2})) = \begin{pmatrix} f(\omega_{k_1}) + O(\frac{1}{n}) & k_1 = k_2 \\ O(\frac{1}{n}) & k_1 \neq k_2, \end{pmatrix}.$$

where

$$f(\omega) = \frac{\sigma^2 |1 + \sum_{j=1}^{q} \theta_j \exp(ij\omega)|^2}{2\pi |1 + \sum_{j=1}^{p} \phi_j \exp(ij\omega)|^2}.$$

In other words, if the time series satisfies an ARMA presentation the DFT is 'near' uncorrelated, its mean is zero and its variance has a well specified parametric form. Using this information we can define a criterion for estimating the parameters. We motivate this criterion through the likelihood, however there are various other methods for motivating the criterion for example the Kullbach-Leibler criterion is an alternative motivation, we comment on this later on.

If the innovations are Gaussian then $\Re J_n(\omega)$ and $\Im J_n(\omega)$ are also Gaussian, thus by using above we approximately have

$$\mathcal{J}_n = \begin{pmatrix} \Re J_n(\omega_1) \\ \Im J_n(\omega_1) \\ \vdots \\ \Re J_n(\omega_{n/2}) \\ \Im J_n(\omega_{n/2}) \end{pmatrix} \sim \mathcal{N}(0, \operatorname{diag}(f(\omega_1), f(\omega_1), \ldots, f(\omega_{n/2}), f(\omega_{n/2}))).$$

In the case that the innovations are not normal then, by Corollary 11.2.1, the above holds asymptotically for a finite number of frequencies. Here we construct the likelihood under normality of the innovations, however, this assumption is <u>not</u> required and is only used to motivate the construction.

Since $\mathcal{J}_n$ is normally distributed random vector with mean zero and 'approximate' diagonal

matrix variance matrix $\text{diag}(f(\omega_1), \ldots, f(\omega_n))$, the negative log-likelihood of $\mathcal{J}_n$ is approximately

$$L_n^w(\boldsymbol{\theta}) = \sum_{k=1}^{n} \left(\log |f_{\boldsymbol{\theta}}(\omega_k)| + \frac{|J_X(\omega_k)|^2}{f_{\boldsymbol{\theta}}(\omega_k)} \right).$$

To estimate the parameter we would choose the $\underline{\theta}$ which minimises the above criterion, that is

$$\widehat{\boldsymbol{\theta}}_n^w = \arg\min_{\boldsymbol{\theta}\in\Theta} L_n^w(\boldsymbol{\theta}), \tag{11.23}$$

where Θ consists of all parameters where the roots of the corresponding characteristic polynomial have absolute value greater than $(1+\delta)$ (note that under this assumption all spectral densities corresponding to these parameters will be bounded away from zero).

Example 11.4.1 *Fitting an ARMA$(1,1)$ model to the data To fit an ARMA model to the data using the Whittle likelihood we use the criterion*

$$L_n^w(\boldsymbol{\theta}) = \sum_{k=1}^{n/2} \left(\log \frac{\sigma^2 |1+\theta e^{i\omega_k}|^2}{2\pi |1-\phi e^{i\omega_k}|} + I_n(\omega_k) \frac{2\pi |1-\phi e^{i\omega_k}|^2}{\sigma^2 |1+\theta e^{i\omega_k}|^2} \right).$$

By differentiating L_n^ω with respect to ϕ, σ^2 and θ we solve these three equations (usually numerically), this gives us the Whittle likelihood estimators.

Whittle (1962) showed that the above criterion is an approximation of the GMLE. The correct proof is quite complicated and uses several matrix approximations due to Grenander and Szegö (1958). Instead we give a heuristic proof which is quite enlightening.

11.4.1 Connecting the Whittle and Gaussian likelihoods

We now give the details of a derivation which explicitly connects the Whittle and the Gaussian likelihood. The details can be found in Subba Rao and Yang (2020). To understand the construction, we review some results from likelihoods of independent random variables and Principal Component analysis.

Background First suppose that $\underline{Y}_n' = (Y_1, \ldots, Y_n)$ is a random vector comprised of independent random variables where $\text{var}[Y_j] = \sigma^2$. Let $\Delta_n = \text{var}[\underline{Y}_n]$. It is easy to show

$$\underline{Y}_n' \Delta_n^{-1} \underline{Y}_n = \sum_{j=1}^{n} \frac{Y_j^2}{\sigma_j^2}.$$

We show below that any random vector can be written in the above form (even when the variance is not a diagonal matrix).

Suppose $\underline{X}_n' = (X_1, \ldots, X_n)$ is a n-dimension random vector with mean zero and variance Σ_n. Let us suppose that Σ_n is non-singular. Since Σ_n is symmetric and positive definite it has the spectral decomposition

$$\Sigma_n = U_n \Delta_n U_n',$$

where $U_n^{-1} = U_n'$ and $\Delta_n = \text{diag}(\lambda_1, \ldots, \lambda_n)$. This means

$$\Sigma_n^{-1} = U_n \Delta_n^{-1} U_n'.$$

The matrix U_n is built of orthogonal vectors but has the additional property that it **decorrelates** the random vector $\underline{X}_n$. To see why, let $U_n = (\underline{u}_1, \ldots, \underline{u}_n)$ define the transformed vector $\underline{Y}_n = U_n' \underline{X}_n$, where the jth entry in the vector is

$$Y_j = \langle \underline{u}_j, \underline{X}_n \rangle = \underline{u}_j' \underline{X}_n = \sum_{s=1}^{n} u_{j,s} X_s.$$

We observe

$$\begin{aligned} \text{var}[\underline{Y}_n] &= \text{var}\left[U_n' \underline{X}_n\right] = U_n' \text{var}[\underline{X}_n] U_n \\ &= U_n' U_n \Delta_n U_n' U_n = \Delta_n, \end{aligned}$$

where we recall $\Delta_n = \text{diag}(\lambda_1, \ldots, \lambda_n)$. This means that the entries of the transformed random vector $\underline{Y}_n' = (Y_1, Y_2, \ldots, Y_n)$ are uncorrelated:

$$\text{cov}\left[Y_{j_1}, Y_{j_2}\right] = 0 \text{ if } j_1 \neq j_2 \text{ and } \text{var}\left[Y_j\right] = \lambda_j \text{ for } j = 1, \ldots, n.$$

Using the inverse transform $\underline{X}_n = (U_n')^{-1} \underline{Y}_n = U_n \underline{Y}_n$ we can represent $\underline{X}_n$ as

$$\underline{X}_n = \sum_{j=1}^{n} Y_j \underline{u}_j.$$

Returning to $\underline{X}_n'\Sigma_n^{-1}\underline{X}_n$ we have

$$\begin{aligned}\underline{X}_n'\Sigma_n^{-1}\underline{X}_n &= \underline{X}_n'U_n\Delta_n^{-1}U_n'\underline{X}_n \\ &= \underline{Y}_n'\Delta_n^{-1}\underline{Y}_n \\ &= \sum_{j=1}^{n}\frac{Y_j^2}{\lambda_j^2}.\end{aligned}$$

However, the above expansion is very specific to the spectral decomposition of Σ_n, which is unique. We now describe an analogous result, which can apply to any unique pairing of matrices.

Biorthogonal transforms and Likelihoods Let us return to the eigenvector matrix U_n. It has the unique property that its vectors $\{\underline{u}_j\}$ are orthogonal and $\text{var}[U_n\underline{X}_n]$ is a diagonal matrix. We now define a more general concept (called a biorthogonal transform) that yields a similar result. For any transform matrix U_n (it does not need to be such that $U_nU_n' = I_n$), there exists a matrix V_n such that $\text{cov}[U_n\underline{X}_n, V_n\underline{X}_n] = \Lambda_n$ (where $\Lambda_n = \text{diag}(\delta_1, \ldots, \delta_n)$ is a diagonal matrix). U_n and V_n are called biorthogonal transforms. By definition of the biorthogonal transform we have

$$\text{cov}\left[U_n\underline{X}_n, V_n\underline{X}_n\right] = U_n\text{var}[\underline{X}_n]V_n^* = \Lambda_n.$$

Rearranging the above gives

$$\text{var}[\underline{X}_n] = U_n^{-1}\Lambda_n(V_n^*)^{-1}$$

next inverting it gives

$$\text{var}[\underline{X}_n]^{-1} = V_n^*\Lambda_n^{-1}U_n.$$

Now we define two transformed random vectors $\underline{Y}_n = U_n\underline{X}_n$ and $\underline{Z}_n = V_n\underline{X}_n$. Let $\underline{Y}_n' = (Y_1, \ldots, Y_n)$ and $\underline{Z}_n' = (Z_1, \ldots, Z_n)$. Then the biorthogonality properties means

$$\text{cov}[Z_{j_1}, Y_{j_2}] = 0 \text{ if } j_1 \neq j_2 \text{ and } \text{cov}[Z_j, Y_j] = \delta_j.$$

Thus

$$\begin{aligned} \underline{X}_n' \Sigma_n^{-1} \underline{X}_n &= \underline{X}_n' V_n^* \Lambda_n^{-1} U_n \underline{X}_n = \underline{Z}_n \Lambda_n^{-1} \underline{Y}_n \\ &= \sum_{j=1}^{n} \frac{Z_n Y_n}{\delta_n}. \end{aligned} \tag{11.24}$$

It is this identity that will link the Gaussian and Whittle likelihood.

The Whittle likelihood We recall that the Whittle likelihood is

$$\mathcal{L}_n(\theta) = \sum_{k=1}^{n} \frac{|J_n(\omega_k)|^2}{f(\omega_k;\theta)}$$

where for simplicity we have ignored the $\log f(\omega_k;\theta)$ term. The DFT vector

$$\underline{Y}_n' = (J_n(\omega_1), \ldots, J_n(\omega_n))'$$

is a linear transform of the observed time series $\underline{X}_n' = (X_1, \ldots, X_n)$. Let F_n be the DFT matrix with $(F_n)_{k,t} = n^{-1/2}\exp(it\omega_k)$, then $\underline{Y}_n = F_n \underline{X}_n$. Then we can write $\mathcal{L}_n(\theta)$ as

$$\mathcal{L}_n(\theta) = \sum_{k=1}^{n} \frac{|(F_n \underline{X}_n)_k|^2}{f(\omega_k;\theta)}.$$

The biorthogonal transform of $\underline{Y}_n$ will give a representation of the Gaussian likelihood in terms of $\underline{Y}_n$. In particular, if V_n is the biorthogonal transform of F_n with respect the variance matrix $\Gamma_n(f_\theta)$ then by using (11.24) we have

$$\underline{X}_n' \Sigma_n(f_\theta)^{-1} \underline{X}_n = \sum_{k=1}^{n} \frac{(F_n^* \underline{X}_n)_k (V_n \underline{X}_n)_k}{\delta_k}.$$

But in Section 10.2.3 we have already encountered the biorthogonal transform to the DFT $J_n(\omega_k)$ it is simply the complete DFT

$$\widetilde{J}_n(\omega; f) = J_n(\omega) + \frac{1}{\sqrt{n}} \sum_{\tau \neq \{1,2,\ldots,n\}} P_X(X_\tau) e^{i\tau\omega} = \frac{1}{\sqrt{n}} \sum_{\tau=-\infty}^{\infty} P_X(X_\tau) e^{i\tau\omega}.$$

where $P_X(X_\tau)$ denotes the projection of X_τ onto $\{X_t\}_{t=1}^n$. Let $\underline{Z}_n' = (\widetilde{J}_n(\omega_1; f), \ldots, \widetilde{J}_n(\omega_n; f))$. By

using the results in Section 10.2.3 we have the biorthogonality

$$\text{cov}[J_n(\omega_{k_1}), \widetilde{J}_n(\omega_{k_2}; f)] = 0 \text{ if } k_1 \neq k_2 \text{ and } \text{cov}[J_n(\omega_k), \widetilde{J}_n(\omega_k; f)] = f(\omega_k) \text{ for } k = 1, \ldots, n.$$

Thus

$$\underline{X}_n' \Sigma_n(f_\theta)^{-1} \underline{X}_n = \sum_{k=1}^{n} \frac{\overline{J_n(\omega_{k_1})} \widetilde{J}_n(\omega_{k_2}; f_\theta)}{f_\theta(\omega_k)}.$$

This places the Gaussian likelihood within the frequency domain and shows that the difference between the Gaussian and Whittle likelihood is

$$\begin{aligned} & \underline{X}_n' \Sigma_n(f_\theta)^{-1} \underline{X}_n - \sum_{k=1}^{n} \frac{|J_n(\omega_k)|^2}{f(\omega_k; \theta)} \\ = & \sum_{k=1}^{n} \frac{\overline{J_n(\omega_{k_1})} \widehat{J}_n(\omega_{k_2}; f_\theta)}{f_\theta(\omega_k)}, \end{aligned}$$

where

$$\widehat{J}_n(\omega; f) = \frac{1}{\sqrt{n}} \sum_{\tau \neq \{1,2,\ldots,n\}} P_X(X_\tau) e^{i\tau\omega}.$$

This result allows us to rewrite the Gaussian likelihood in the frequency domain and understand how precisely the two likelihoods are connected.

11.4.2 Sampling properties of the Whittle likelihood estimator

Lemma 11.4.1 (Consistency) *Suppose that $\{X_t\}$ is a causal ARMA process with parameters $\boldsymbol{\theta}$ whose roots lie outside the $(1+\delta)$-circle (where $\delta > 0$ is arbitrary). Let $\widehat{\boldsymbol{\theta}}^w$ be defined as in (11.23) and suppose that $\mathrm{E}(\varepsilon_t^4) < \infty$. Then we have*

$$\widehat{\boldsymbol{\theta}}^w \xrightarrow{\mathcal{P}} \boldsymbol{\theta}.$$

PROOF. To show consistency we need to show pointwise convergence and equicontinuity of $\frac{1}{n}\mathcal{L}_n$. Let

$$L^w(\boldsymbol{\theta}) = \frac{1}{2\pi} \int_0^{2\pi} \left(\log f_{\boldsymbol{\theta}}(\omega) + \frac{f_{\boldsymbol{\theta}_0}(\omega)}{f_{\boldsymbol{\theta}}(\omega)} \right) d\omega.$$

It is straightforward to show that $\mathrm{E}(\frac{1}{n}\mathcal{L}_n^w(\boldsymbol{\theta})) \to \tilde{\mathcal{L}}_n(\boldsymbol{\theta})$. Next we evaluate the variance, to do this we use Proposition 10.7.1 and obtain

$$\mathrm{var}\left[\frac{1}{n}L_n^w(\boldsymbol{\theta})\right] = \frac{1}{n^2}\sum_{k_1,k_2=1}^{n}\frac{1}{f_{\boldsymbol{\theta}}(\omega_{k_1})f_{\boldsymbol{\theta}}(\omega_{k_2})}\mathrm{cov}(|J_n(\omega_{k_1})|^2,|J_n(\omega_{k_2})|^2) = O(\frac{1}{n}).$$

Thus we have

$$\frac{1}{n}L_n^w(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} L^w(\boldsymbol{\theta}).$$

To show equicontinuity we apply the mean value theorem to $\frac{1}{n}L_n^w$. We note that because the parameters $(\underline{\phi},\underline{\theta}) \in \Theta$, have characteristic polynomial whose roots are greater than $(1+\delta)$ then $f_{\boldsymbol{\theta}}(\omega)$ is bounded away from zero (there exists a $\delta^* > 0$ where $\inf_{\omega,\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega) \geq \delta^*$). Hence it can be shown that there exists a random sequence $\{\mathcal{K}_n\}$ such that $|\frac{1}{n}L_n^w(\boldsymbol{\theta}_1) - \frac{1}{n}L_n^w(\boldsymbol{\theta}_2))| \leq \mathcal{K}_n(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|)$ and $\mathcal{K}_n$ converges almost surely to a finite constant as $n \to \infty$. Therefore $\frac{1}{n}L_n$ is stochastically equicontinuous. Since the parameter space Θ is compact, the three standard conditions are satisfied and we have consistency of the Whittle estimator. □

To show asymptotic normality we note that $\frac{1}{n}L_n^w(\boldsymbol{\theta})$ can be written as a quadratic form

$$\frac{1}{n}L_n^w(\boldsymbol{\theta}) = \int_0^{2\pi}\log f_{\boldsymbol{\theta}}(\omega_k) + \frac{1}{n}\sum_{r=-(n-1)}^{n-1} d_n(r;\boldsymbol{\theta})\sum_{k=1}^{n-|r|} X_k X_{k+r}$$

where

$$d_n(r;\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^{n} f_{\boldsymbol{\theta}}(\omega_k)^{-1}\exp(ir\omega_k).$$

Using the above quadratic form and it's derivatives wrt $\boldsymbol{\theta}$ one can show normality of the Whittle likelihood under various dependence conditions on the time series. Using this result, in the following theorem we show asymptotic normality of the Whittle estimator. Note, this result not only applies to linear time series, but several types of nonlinear time series too.

Theorem 11.4.1 *Let us suppose that $\{X_t\}$ is a strictly stationary time series with a sufficient dependence structure (such as linearity, mixing at a certain rate, etc.) with spectral density function*

$f_{\boldsymbol{\theta}}(\omega)$ *and* $\mathrm{E}|X_t^4| < \infty$. *Let*

$$L_n^w(\boldsymbol{\theta}) = \sum_{k=1}^{n} \Big(\log |f_{\boldsymbol{\theta}}(\omega_k)| + \frac{|J_n(\omega_k)|^2}{f_{\boldsymbol{\theta}}(\omega_k)} \Big),$$

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} L_n^w(\boldsymbol{\theta}) \qquad \boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}\in\Theta} L^w(\boldsymbol{\theta})$$

Then we have

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, 2V^{-1} + V^{-1}WV^{-1})$$

where

$$\begin{aligned} V &= \frac{1}{2\pi}\int_0^{2\pi} \Big(\frac{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega)}{f_{\boldsymbol{\theta}}(\omega)}\Big)\Big(\frac{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega)}{f_{\boldsymbol{\theta}}(\omega)}\Big)' d\omega \\ W &= \frac{2}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi} \big(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega_1)^{-1}\big)\big(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega_2)^{-1}\big)' f_{4,\boldsymbol{\theta}_0}(\omega_1, -\omega_1, \omega_2) d\omega_1 d\omega_2, \end{aligned}$$

and $f_{4,\boldsymbol{\theta}_0}(\omega_1,\omega_2,\omega_3)$ *is the fourth order spectrum of* $\{X_t\}$.

We now apply the above result to the case of linear time series. We now show that in this case, in the fourth order cumulant term, W, falls out. This is due to the following lemma.

Lemma 11.4.2 *Suppose that the spectral density has the form* $f(\omega) = \sigma^2|1 + \sum_{j=1}^{\infty}\psi_j \exp(ij\omega)|^2$ *and* $\inf f(\omega) > 0$. *Then we have*

$$\frac{1}{2\pi}\int_0^{2\pi} \log f(\omega) d\omega = \log \sigma^2$$

PROOF. Since $f(z)$ is non-zero for $|z| \leq 1$, then $\log f(z)$ has no poles in $\{z; |z| \leq 1\}$. Thus we have

$$\begin{aligned} &\frac{1}{2\pi}\int_0^{2\pi} \log f(\omega) d\omega = \frac{1}{2\pi}\int_0^{2\pi} \log \sigma^2 d\omega + \frac{1}{2\pi}\int_0^{2\pi} \log |1 + \sum_{j=1}^{\infty}\psi_j \exp(ij\omega)|^2 d\omega \\ = &\frac{1}{2\pi}\int_0^{2\pi} \log \sigma^2 d\omega + \frac{1}{2\pi}\int_{|z|=1} \log |1 + \sum_{j=1}^{\infty}\psi_j z|^2 dz \\ = &\frac{1}{2\pi}\int_0^{2\pi} \log \sigma^2 d\omega. \end{aligned}$$

An alternative proof is that since $f(z)$ is analytic and does not have any poles for $|z| \leq 1$, then

$\log f(z)$ is also analytic in the region $|z| \leq 1$, thus for $|z| \leq 1$ we have the power series expansion $\log|1+\sum_{j=1}^{\infty}\psi_j \exp(ij\omega)|^2 = \sum_{j=1}^{\infty} b_j z^j$ (a Taylor expansion about $\log 1$). Using this we have

$$\begin{aligned} & \frac{1}{2\pi}\int_0^{2\pi} \log|1+\sum_{j=1}^{\infty}\psi_j \exp(ij\omega)|^2 d\omega = \frac{1}{2\pi}\int_0^{2\pi}\sum_{j=1}^{\infty} b_j \exp(ij\omega)d\omega \\ = & \frac{1}{2\pi}\sum_{j=1}^{\infty} b_j \int_0^{2\pi}\exp(ij\omega)d\omega = 0, \end{aligned}$$

and we obtain the desired result. □

Lemma 11.4.3 *Suppose that $\{X_t\}$ is a linear ARMA time series $X_t - \sum_{j=1}^{p}\phi_j X_{t-j} = \sum_{i=1}^{q}\theta_i\varepsilon_{t-i} + \varepsilon_t$, where $\mathrm{E}[\varepsilon_t]=0$, $\mathrm{var}[\varepsilon_t]=\sigma^2$ and $\mathrm{E}[\varepsilon_t^4]<\infty$. Let $\boldsymbol{\theta} = (\{\phi_j,\theta_j\})$, then we have $W=0$ and*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n^w - \boldsymbol{\theta}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2V^{-1}).$$

PROOF. The result follows from Theorem 11.4.1, however we need to show that in the case of linearity that $W=0$.

We use Example 10.7.1 for linear processes to give $f_{4,\boldsymbol{\theta}}(\omega_1,\omega_1,-\omega_2) = \kappa_4|A(\omega_1)|^2|A(\omega_2)|^2 = \frac{\kappa_4}{\sigma^4}f(\omega_1)f(\omega_2)$. Substituting this into W gives

$$\begin{aligned} W &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{2\pi}\left(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega_1)^{-1}\right)\left(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega_2)^{-1}\right)' f_{4,\boldsymbol{\theta}_0}(\omega_1,-\omega_1,\omega_2)d\omega_1 d\omega_2 \\ &= \frac{\kappa_4}{\sigma^4}\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega)}{f_{\boldsymbol{\theta}}(\omega)^2} f_{\boldsymbol{\theta}}(\omega)d\omega\right)^2 = \frac{\kappa_4}{\sigma^4}\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega)}{f_{\boldsymbol{\theta}}(\omega)}d\omega\right)^2 \\ &= \frac{\kappa_4}{\sigma^4}\left(\frac{1}{2\pi}\int_0^{2\pi}\nabla_{\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\omega)d\omega\right)^2 \\ &= \frac{\kappa_4}{\sigma^4}\left(\frac{1}{2\pi}\nabla_{\boldsymbol{\theta}}\int_0^{2\pi}\log f_{\boldsymbol{\theta}}(\omega)d\omega\right)^2 = \frac{\kappa_4}{\sigma^4}\left(\nabla_{\boldsymbol{\theta}}\log\frac{\sigma^2}{2\pi}\right)^2 = 0, \end{aligned}$$

where by using Lemma 11.4.2 we have $\int_0^{2\pi}\log f_{\boldsymbol{\theta}}(\omega)d\omega = 2\pi\log\frac{\sigma^2}{2\pi}$ and since $\boldsymbol{\theta}$ does not include σ^2 we obtain the above. Hence for linear processes the higher order cumulant does not play an asymptotic role in the variance thus giving the result. □

On first appearances there does not seem to be a connection between the Whittle likelihood and the sample autocorrelation estimator defined in Section 8.2.1. However, we observe that the variance of both estimators, under linearity, do not contain the fourth order cumulant (even for non-Gaussian linear time series). In Section 11.5 we explain there is a connection between the two,

and it is this connection that explains away this fourth order cumulant term.

Remark 11.4.1 *Under linearity, the GMLE and the Whittle likelihood are asymptotically equivalent, therefore they have the same asymptotic distributions. The GMLE has the asymptotic distribution* $\sqrt{n}(\hat{\underline{\phi}}_n - \underline{\phi}, \hat{\underline{\theta}}_n - \underline{\theta}) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, \Lambda^{-1})$*, where*

$$\Lambda = \begin{pmatrix} \mathrm{E}(U_t U_t') & \mathrm{E}(V_t U_t') \\ \mathrm{E}(U_t V_t') & \mathrm{E}(V_t V_t') \end{pmatrix}$$

and $\{U_t\}$ *and* $\{V_t\}$ *are autoregressive processes which satisfy* $\phi(B)U_t = \varepsilon_t$ *and* $\theta(B)V_t = \varepsilon_t$*.*

By using the similar derivatives to those given in (9.26) we can show that

$$\begin{pmatrix} \mathrm{E}(U_t U_t') & \mathrm{E}(V_t U_t') \\ \mathrm{E}(U_t V_t') & \mathrm{E}(V_t V_t') \end{pmatrix} = \frac{1}{2\pi} \int_0^{2\pi} \left(\frac{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega)}{f_{\boldsymbol{\theta}}(\omega)} \right) \left(\frac{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega)}{f_{\boldsymbol{\theta}}(\omega)} \right)' d\omega.$$

11.5 Ratio statistics in Time Series

We recall from (11.4) that the covariance can be written as a general periodogram mean which has the form

$$A(\phi, I_n) = \frac{1}{n} \sum_{k=1}^{n} I_n(\omega_k)\phi(\omega_k). \tag{11.25}$$

The variance of this statistic is

$$\begin{aligned} \mathrm{var}(A(\phi, I_n)) &= \frac{1}{n^2} \sum_{k_1,k_2=1}^{n} \phi(\omega_{k_1})\overline{\phi(\omega_{k_1})}\mathrm{cov}(|J_n(\omega_{k_1})|^2, |J_n(\omega_{k_2})|^2) \\ &= \frac{1}{n^2} \sum_{k_1,k_2=1}^{n} \phi(\omega_{k_1})\overline{\phi(\omega_{k_1})}\Big[\mathrm{cov}(J_n(\omega_{k_1}), J_n(\omega_{k_2}))\mathrm{cov}(\overline{J_n(\omega_{k_1})}, \overline{J_n(\omega_{k_2})}) \\ &\quad + \mathrm{cov}(J_n(\omega_{k_1}), \overline{J_n(\omega_{k_2})})\mathrm{cov}(\overline{J_n(\omega_{k_1})}, J_n(\omega_{k_2})) \\ &\quad + cum(J_n(\omega_{k_1}), \overline{J_n(\omega_{k_2})}, J_n(\omega_{k_2}), \overline{J_n(\omega_{k_2})})\Big]. \end{aligned} \tag{11.26}$$

By using Proposition 10.7.1 we have

$$\begin{aligned}
&\mathrm{cov}(|J_n(\omega_{k_1})|^2, |J_n(\omega_{k_2})|^2) = \\
&\left[f(\omega_{k_1})I(k_1 = k_2) + O\left(\frac{1}{n}\right)\right]^2 + \left[f(\omega_{k_1})I(k_1 = n - k_2) + O\left(\frac{1}{n}\right)\right]\left[f(\omega_{k_1})I(n - k_1 = k_2) + O\left(\frac{1}{n}\right)\right] \\
&+\frac{1}{n}f_4(\omega_1, -\omega_1, \omega_2) + O\left(\frac{1}{n^2}\right). \qquad (11.27)
\end{aligned}$$

Substituting (11.27) into (11.26) the above gives

$$\begin{aligned}
&\mathrm{var}(A(\phi, I_n)) \\
= \;& \frac{1}{n^2}\sum_{k=1}^{n}|\phi(\omega_k)|^2 f(\omega_k)^2 + \frac{1}{n^2}\sum_{k=1}^{n}\phi(\omega_k)\overline{\phi(\omega_{n-k})}f(\omega_k)^2 \\
&+\frac{1}{n^3}\sum_{k_1,k_2=1}^{n}\phi(\omega_{k_1})\overline{\phi(\omega_{k_2})}f_4(\omega_{k_1}, -\omega_{k_1}, \omega_{k_2}) + O(\frac{1}{n^2}) \\
= \;& \frac{1}{n}\int_0^{2\pi}|\phi(\omega)|^2 f(\omega)^2 d\omega + \frac{1}{n}\int_0^{2\pi}\phi(\omega)\overline{\phi(2\pi-\omega)}f(\omega)^2 d\omega \\
&+\frac{1}{n}\int_0^{2\pi}\int_0^{2\pi}\phi(\omega_1)\overline{\phi(\omega_2)}f_4(\omega_1, -\omega_1, \omega_2)d\omega_1 d\omega_2 + O(\frac{1}{n^2}), \qquad (11.28)
\end{aligned}$$

where f_4 is the fourth order cumulant of $\{X_t\}$. From above we see that unless ϕ satisfies some special conditions, $\mathrm{var}(A(\phi, I_n))$ contains the fourth order spectrum, which can be difficult to estimate. There are bootstrap methods which can be used to estimate the variance or finite sample distribution, but simple bootstrap methods, such as the frequency domain bootstrap, cannot be applied to $A(\phi, I_n)$, since it is unable to capture the fourth order cumulant structure. However, in special cases the fourth order structure is disappears, we consider this case below and then discuss how this case can be generalised.

Lemma 11.5.1 *Suppose $\{X_t\}$ is a linear time series, with spectral density $f(\omega)$. Let $A(\phi, I_n)$ be defined as in (11.25) and suppose the condition*

$$A(\phi, f) = \int \phi(\omega)f(\omega)d\omega = 0 \qquad (11.29)$$

holds, then

$$\mathrm{var}(A(\phi, I_n)) = \frac{1}{n}\int_0^{2\pi}|\phi(\omega)|^2 f(\omega)^2 d\omega + \frac{1}{n}\int_0^{2\pi}\phi(\omega)\overline{\phi(2\pi-\omega)}f(\omega)^2 d\omega.$$

PROOF. By using (11.28) we have

$$\begin{aligned} & \operatorname{var}(A(\phi, I_n)) \\ = & \frac{1}{n}\int_0^{2\pi} |\phi(\omega)|^2 f(\omega)^2 d\omega + \frac{1}{n}\int_0^{2\pi} \phi(\omega)\overline{\phi(2\pi-\omega)} f(\omega)^2 d\omega \\ & \frac{1}{n}\int_0^{2\pi}\int_0^{2\pi} \phi(\omega_1)\overline{\phi(\omega_2)} f_4(\omega_1, -\omega_1, \omega_2) + O(\frac{1}{n^2}). \end{aligned}$$

But under linearity $f_4(\omega_1, -\omega_1, \omega_2) = \frac{\kappa_4}{\sigma^4} f(\omega_1) f(\omega_2)$, substituting this into the above gives

$$\begin{aligned} & \operatorname{var}(A(\phi, I_n)) \\ = & \frac{1}{n}\int_0^{2\pi} |\phi(\omega)|^2 f(\omega)^2 d\omega + \frac{1}{n}\int_0^{2\pi} \phi(\omega)\overline{\phi(2\pi-\omega)} f(\omega)^2 d\omega \\ & \frac{\kappa_4}{\sigma^4}\frac{1}{n}\int_0^{2\pi}\int_0^{2\pi} \phi(\omega_1)\overline{\phi(\omega_2)} f(\omega_1) f(\omega_2) d\omega_1 d\omega_2 + O\left(\frac{1}{n^2}\right) \\ = & \frac{1}{n}\int_0^{2\pi} |\phi(\omega)|^2 f(\omega)^2 d\omega + \frac{1}{n}\int_0^{2\pi} \phi(\omega)\overline{\phi(2\pi-\omega)} f(\omega)^2 d\omega \\ & + \frac{\kappa_4}{\sigma^4}\frac{1}{n}\left|\int_0^{2\pi} \phi(\omega) f(\omega) d\omega\right|^2 + O\left(\frac{1}{n^2}\right). \end{aligned}$$

Since $\int \phi(\omega) f(\omega) d\omega = 0$ we have the desired result. □

Example 11.5.1 (The Whittle likelihood) *Let us return to the Whittle likelihood in the case of linearity. In Lemma 11.4.3 we showed that the fourth order cumulant term does not play a role in the variance of the ARMA estimator. We now show that condition (11.29) holds.*

Consider the partial derivative of the Whittle likelihood

$$\nabla_{\boldsymbol{\theta}} L_n^w(\boldsymbol{\theta}) = \sum_{k=1}^{n}\left(\frac{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)} - \frac{I_n(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)^2}\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega_k)\right).$$

To show normality we consider the above at the true parameter $\boldsymbol{\theta}$, this gives

$$\nabla_{\boldsymbol{\theta}} L_n^w(\boldsymbol{\theta}) = \sum_{k=1}^{n}\left(\frac{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)} - \frac{I_n(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)^2}\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega_k)\right).$$

Only the second term of the above is random, therefore it is only this term that yields the variance. Let

$$A(f_{\boldsymbol{\theta}}^{-2}\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}, I_n) = \frac{1}{n}\sum_{k=1}^{n}\frac{I_n(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)^2}\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega_k).$$

To see whether this term satisfies the conditions of Lemma 11.5.1 we evaluate

$$\begin{aligned} A(f_{\boldsymbol{\theta}}^{-2}\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}}) &= \int_0^{2\pi} \frac{f_{\boldsymbol{\theta}}(\omega)}{f_{\boldsymbol{\theta}}(\omega)^2}\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\omega) \\ &= \int_0^{2\pi} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\omega) \\ &= \nabla_{\boldsymbol{\theta}} \int_0^{2\pi} \log f_{\boldsymbol{\theta}}(\omega) = \nabla_{\boldsymbol{\theta}} \frac{1}{2\pi}\int_0^{2\pi} \log f_{\boldsymbol{\theta}}(\omega) d\omega = 0, \end{aligned}$$

by using Lemma 11.4.2. Thus we see that the derivative of the Whittle likelihood satisfies the condition (11.29). Therefore the zero cumulant term is really due to this property. □

The Whittle likelihood is a rather special example. However we now show that any statistic of the form $A(\phi, I_n)$ can be transformed such that the resulting transformed statistic satisfies condition (11.29). To find the suitable transformation we recall from Section 8.2.1 that the variance of $\widehat{c}_n(r)$ involves the fourth order cumulant, but under linearity the sample correlation $\widehat{\rho}_n(r) = \widehat{c}_n(r)/\widehat{c}_n(0)$ does given not. Returning to the frequency representation of the autocovariance given in (11.5) we observe that

$$\widehat{\rho}_n(r) = \frac{1}{\widehat{c}_n(0)}\frac{1}{n}\sum_{k=1}^{n/2} I_n(\omega_k)\exp(ir\omega_k) \approx \frac{1}{\widehat{c}_n(0)}\frac{1}{n}\sum_{k=1}^{n} I_n(\omega_k)\exp(ir\omega_k),$$

(it does not matter whether we sum over n or $n/2$ for the remainder of this section we choose the case of summing over n). Motivated by this example we define the so called 'ratio' statistic

$$\widetilde{A}(\phi, I_n) = \frac{1}{n}\sum_{k=1}^{n}\frac{I_n(\omega_k)\phi(\omega_k)}{\widehat{c}_n(0)} = \frac{1}{n}\sum_{k=1}^{n}\frac{I_n(\omega_k)\phi(\omega_k)}{\hat{F}_n(2\pi)}, \tag{11.30}$$

where $\hat{F}_n(2\pi) = \frac{1}{n}\sum_{k=1}^{n} I_n(\omega_k) = \frac{1}{n}\sum_{t=1}^{n} X_t^2 = \widehat{c}_n(0)$. We show in the following lemma that $\widetilde{A}(\phi, I_n)$ can be written in a form that 'almost' satisfies condition (11.29).

Lemma 11.5.2 *Let us suppose that $\widetilde{A}(\phi, I_n)$ satisfies (11.30) and*

$$\widetilde{A}(\phi, f) = \frac{1}{n}\sum_{k=1}^{n}\frac{f(\omega_k)\phi(\omega_k)}{F_n(2\pi)},$$

where $F_n(2\pi) = \frac{1}{n}\sum_{j=1}^n f(\omega_k)$. *Then we can represent* $\tilde{A}(\phi, I_n)$ *as*

$$\widetilde{A}(\phi, I_n) - \widetilde{A}(\phi, f) = \frac{1}{F(2\pi)\hat{F}_n(2\pi)}\frac{1}{n}\sum_{k=1}^n \psi_n(\omega_k)I_n(\omega_k),$$

where

$$\psi_n(\omega_k) = \phi(\omega_k)F_n(2\pi) - \frac{1}{n}\sum_{j=1}^n \phi(\omega_j)f(\omega_j) \qquad \textit{and} \qquad \frac{1}{n}\sum_{k=1}^n \psi(\omega_k)f(\omega_k) = 0. \tag{11.31}$$

PROOF. Basic algebra gives

$$\begin{aligned}
\tilde{A}(\phi, I_n) - \tilde{A}(\phi, f) &= \frac{1}{n}\sum_{k=1}^n \left(\frac{\phi(\omega_k)I_n(\omega_k)}{\hat{F}_n(2\pi)} - \frac{\phi(\omega_k)f(\omega_k)}{F_n(2\pi)}\right) \\
&= \frac{1}{n}\sum_{k=1}^n \left(\frac{\phi(\omega_k)F_n(2\pi)I_n(\omega_k) - \phi(\omega_k)\hat{F}_n(2\pi)f(\omega_k)}{F_n(2\pi)\hat{F}_n(2\pi)}\right) \\
&= \frac{1}{n}\sum_{k=1}^n \left(\phi(\omega_k)F_n(2\pi) - \frac{1}{n}\sum_{k=1}^n \phi(\omega_k)f(\omega_k)\right)\frac{I_n(\omega_k)}{F_n(2\pi)\hat{F}_n(2\pi)} \\
&= \frac{1}{n}\sum_{k=1}^n \frac{\psi(\omega_k)I_n(\omega_k)}{F_n(2\pi)\hat{F}_n(2\pi)},
\end{aligned}$$

where $F_n(2\pi)$ and ψ are defined as above. To show (11.31), again we use basic algebra to give

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^n \psi(\omega_k)f(\omega_k) &= \frac{1}{n}\sum_{k=1}^n \left(\phi(\omega)F_n(2\pi) - \frac{1}{n}\sum_{j=1}^n \phi(\omega_j)f(\omega_j)\right)f(\omega_k) \\
&= \frac{1}{n}\sum_{k=1}^n \phi(\omega_k)f(\omega_k)F_n(2\pi) - \frac{1}{n}\sum_{k=1}^n \phi(\omega_k)f(\omega_k)\frac{1}{n}\sum_{j=1}^n f(\omega_j) = 0.
\end{aligned}$$

□

From the lemma above we see that $\tilde{A}(\phi, I_n) - \tilde{A}(\phi, f)$ almost seems to satisfy the conditions in Lemma 11.5.1, the only difference is the random term $\widehat{c}_n(0) = \widehat{F}_n(2\pi)$ in the denominator. We now show that that we can replace $\widehat{F}_n(2\pi)$ with it's limit and that error is asymptotically negligible. Let

$$\tilde{A}(\phi, I_n) - \tilde{A}(\phi, f) = \frac{1}{F_n(2\pi)\hat{F}_n(2\pi)}\frac{1}{n}\sum_{k=1}^n \psi_n(\omega_k)I_n(\omega_k) := \widetilde{B}(\psi, I_n)$$

and

$$B(\psi_n, I_n) = \frac{1}{F_n(2\pi)^2}\frac{1}{n}\sum_{k=1}^{n}\psi(\omega_k)I_n(\omega_k).$$

By using the mean value theorem (basically the Delta method) and expanding $\widetilde{B}(\psi_n, I_n)$ about $B(\psi_n, I_n)$ (noting that $B(\phi_n, f) = 0$) gives

$$\begin{aligned}
& \widetilde{B}(\psi, I_n) - B(\psi, I_n) \\
= & \underbrace{\left(\hat{F}_n(2\pi) - F_n(2\pi)\right)}_{O_p(n^{-1/2})}\frac{1}{\overline{F_n(2\pi)}^3}\underbrace{\frac{1}{n}\sum_{k=1}^{n}\psi_n(\omega_k)I_n(\omega_k)}_{O_p(n^{-1/2})} = O_p(\frac{1}{n}),
\end{aligned}$$

where $\overline{F}_n(2\pi)$ lies between $F_n(2\pi)$ and $\widehat{F}_n(2\pi)$. Therefore the limiting distribution variance of $\tilde{A}(\phi, I_n) - \tilde{A}(\phi, f)$ is determined by

$$\widetilde{A}(\phi, I_n) - \widetilde{A}(\phi, f) = B(\psi_n, I_n) + O_p(n^{-1/2}).$$

$B(\psi_n, I_n)$ does satisfy the conditions in (11.29) and the lemma below immediately follows.

Lemma 11.5.3 *Suppose that $\{X_t\}$ is a linear time series, then*

$$\mathrm{var}(B(\psi_n, I_n)) = \frac{1}{n}\int_0^{2\pi}|\psi(\omega)|^2 f(\omega)^2 d\omega + \frac{1}{n}\int_0^{2\pi}\psi(\omega)\overline{\psi(2\pi-\omega)}f(\omega)^2 d\omega + O(\frac{1}{n^2}),$$

where

$$\psi(\omega) = \phi(\omega)F(2\pi) - \frac{1}{2\pi}\int_0^{2\pi}\phi(\omega)f(\omega)d\omega.$$

Therefore, the limiting variance of $\widetilde{A}(\phi, I_n)$ is

$$\frac{1}{n}\int_0^{2\pi}|\psi(\omega)|^2 f(\omega)^2 d\omega + \frac{1}{n}\int_0^{2\pi}\psi(\omega)\overline{\psi(2\pi-\omega)}f(\omega)^2 d\omega + O(\frac{1}{n^2}).$$

This is a more elegant explanation as to why under linearity the limiting variance of the correlation estimator does not contain the fourth order cumulant term. It also allows for a general class of statistics.

Remark 11.5.1 (Applications) *As we remarked above, many statistics can be written as a ratio statistic. The advantage of this is that the variance of the limiting distribution is only in terms of the spectral densities, and not any other higher order terms (which are difficult to estimate). Another perk is that simple schemes such as the frequency domain bootstrap can be used to estimate the finite sample distributions of statistics which satisfy the assumptions in Lemma 11.5.1 or is a ratio statistic (so long as the underlying process is linear), see Dahlhaus and Janas (1996) for the details. The frequency domain bootstrap works by constructing the DFT from the data $\{J_n(\omega)\}$ and dividing by the square root of either the nonparametric estimator of f or a parametric estimator, ie. $\{J_n(\omega)/\sqrt{\hat{f}_n(\omega)}\}$, these are close to constant variance random variables. $\{\hat{J}_\varepsilon(\omega_k) = J_n(\omega_k)/\sqrt{\hat{f}_n(\omega_k)}\}$ is bootstrapped, thus $J_n^*(\omega_k) = \hat{J}_\varepsilon^*(\omega_k)\sqrt{\hat{f}_n(\omega_k)}$ is used as the bootstrap DFT. This is used to construct the bootstrap estimator, for example*

- *The Whittle likelihood estimator.*
- *The sample correlation.*

With these bootstrap estimators we can construct an estimator of the finite sample distribution.

The nature of frequency domain bootstrap means that the higher order dependence structure is destroyed, eg. $cum^*(J_n^*(\omega_{k_1}), J_n^*(\omega_{k_2}), \ldots, J_n^*(\omega_{k_r})) = 0$ (where cum^* is the cumulant with respect to the bootstrap measure) if all the k_is that are not the same. However, we know from Proposition 10.7.1 that for the actual DFT this is not the case, there is still some 'small' dependence, which can add up. Therefore, the frequency domain bootstrap is unable to capture any structure beyond the second order. This means for a linear time series which is not Gaussian the frequency domain bootstrap cannot approximate the distribution of the sample covariance (since it is asymptotically with normal with a variance which contains the forth order cumulant), but it can approximate the finite sample distribution of the correlation.

Remark 11.5.2 (Estimating κ_4 in the case of linearity) *Suppose that $\{X_t\}$ is a linear time series*

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

with $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{var}(\varepsilon_t) = \sigma^2$ and $cum_4(\varepsilon_t) = \kappa_4$. Then we can use the spectral density estimator to estimate κ_4 without any additional assumptions on $\{X_t\}$ (besides linearity). Let $f(\omega)$ denote the

spectral density of $\{X_t\}$ *and* $g_2(\omega)$ *the spectral density of* $\{X_t^2\}$*, then it can be shown that*

$$\kappa_4 = \frac{2\pi g_2(0) - 4\pi \int_0^{2\pi} f(\omega)^2 d\omega}{\big(\int_0^{2\pi} f(\omega) d\omega\big)^2}.$$

Thus by estimating f *and* g_2 *we can estimate* κ_4*.*

Alternatively, we can use the fact that for linear time series, the fourth order spectral density $f_4(\omega_1, \omega_2, \omega_3) = \kappa_4 A(\omega_1)A(\omega_2)A(\omega_3)A(-\omega_1 - \omega_2 - \omega_3)$*. Thus we have*

$$\kappa_4 = \frac{\sigma^4 f_4(\omega_1, -\omega_1, \omega_2)}{f(\omega_1)f(\omega_2)}.$$

This just demonstrates, there is no unique way to solve a statistical problem!

11.6 Goodness of fit tests for linear time series models

As with many other areas in statistics, we often want to test the appropriateness of a model. In this section we briefly consider methods for validating whether, say an ARMA(p, q), is the appropriate model to fit to a time series. One method is to fit the model to the data and the estimate the residuals and conduct a Portmanteau test (see Section 6, equation (8.15)) on the estimated residuals. It can be shown that if model fitted to the data is the correct one, the estimated residuals behave almost like the true residuals in the model and the Portmanteau test statistic

$$\mathcal{S}_h = n \sum_{r=1}^{h} |\hat{\rho}_n(r)|^2,$$

where $\hat{\rho}_n(r) = \hat{c}_n(r)/\hat{c}_n(0)$

$$\hat{c}_n(r) = \frac{1}{n} \sum_{t=1}^{n-|r|} \hat{\varepsilon}_t \hat{\varepsilon}_{t+r}$$

should be asymptotically a chi-squared. An alternative (but somehow equivalent) way to do the test, is through the DFTs. We recall if the time series is linear then (11.11) is true, thus

$$\frac{I_X(\omega)}{f_{\boldsymbol{\theta}}(\omega)} = |J_\varepsilon(\omega)|^2 + o_p(1).$$

Therefore, if we fit the correct model to the data we would expect that

$$\frac{I_X(\omega)}{f_{\hat{\boldsymbol{\theta}}}(\omega)} = |J_\varepsilon(\omega)|^2 + o_p(1).$$

where $\hat{\boldsymbol{\theta}}$ are the model parameter estimators. Now $|J_\varepsilon(\omega)|^2$ has the special property that not only is it almost uncorrelated at various frequencies, but it is constant over all the frequencies. Therefore, we would expect that

$$\frac{1}{2\pi\sqrt{n}} \sum_{k=1}^{n/2} \big(\frac{I_X(\omega)}{f_{\hat{\boldsymbol{\theta}}}(\omega)} - 2\big) \xrightarrow{\mathcal{D}} N(0,1)$$

Thus, as an alternative to the goodness fit test based on the portmanteau test statistic we can use the above as a test statistic, noting that under the alternative the mean would be different.

11.7 Appendix

Returning the the Gaussian likelihood for the ARMA process, defined in (9.25), we rewrite it as

$$L_n(\boldsymbol{\theta}) = -\left(\det|R_n(\boldsymbol{\theta})| + \mathbf{X}_n' R_n(\boldsymbol{\theta})^{-1}\mathbf{X}_n\right) = -\left(\det|R_n(f_{\boldsymbol{\theta}})| + \mathbf{X}_n' R_n(f_{\boldsymbol{\theta}})^{-1}\mathbf{X}_n\right), \qquad (11.32)$$

where $R_n(f_{\boldsymbol{\theta}})_{s,t} = \int f_{\boldsymbol{\theta}}(\omega)\exp(i(s-t)\omega)d\omega$ and $\mathbf{X}_n' = (X_1,\ldots,X_n)$. We now show that $L_n(\boldsymbol{\theta}) \approx -L_n^w(\boldsymbol{\theta})$.

Lemma 11.7.1 *Suppose that $\{X_t\}$ is a stationary ARMA time series with absolutely summable covariances and $f_{\boldsymbol{\theta}}(\omega)$ is the corresponding spectral density function. Then*

$$\det|R_n(f_{\boldsymbol{\theta}})| + \mathbf{X}_n' R_n(f_{\boldsymbol{\theta}})^{-1}\mathbf{X}_n = \sum_{k=1}^{n}\left(\log|f_{\boldsymbol{\theta}}(\omega_k)| + \frac{|J_n(\omega_k)|^2}{f_{\boldsymbol{\theta}}(\omega_k)}\right) + O(1),$$

for large n.

PROOF. There are various ways to precisely prove this result. All of them show that the Toeplitz matrix can in some sense be approximated by a circulant matrix. This result uses Szegö's identity (Grenander and Szegö (1958)). The main difficulty in the proof is showing that $R_n(f_{\boldsymbol{\theta}})^{-1} \approx U_n(f_{\boldsymbol{\theta}}^{-1})$, where $U_n(f_{\boldsymbol{\theta}}^{-1})_{s,t} = \int f_{\boldsymbol{\theta}}(\omega)^{-1}\exp(i(s-t)\omega)d\omega$. An interesting derivation is given in Brockwell and Davis (1998), Section 10.8. The main ingredients in the proof are:

1. For a sufficiently large m, $R_n(f_{\boldsymbol{\theta}})^{-1}$ can be approximated by $R_n(g_m)^{-1}$, where g_m is the spectral density of an mth order autoregressive process (this follows from Lemma 10.6.2), and showing that

$$\begin{aligned}\underline{X}_n' R_n(f_{\boldsymbol{\theta}})^{-1}\underline{X}_n - \underline{X}_n' R_n(g_m)^{-1}\underline{X}_n &= \underline{X}_n' \left[R_n(f_{\boldsymbol{\theta}})^{-1} - R_n(g_m)^{-1}\right]\underline{X}_n \\ &= \underline{X}_n' R_n(g_m)^{-1}\left[R_n(g_m) - R_n(f_{\boldsymbol{\theta}})\right] R_n(f_{\boldsymbol{\theta}}^{-1})\underline{X}_n \to 0.\end{aligned}$$

2. From Section 6.3, we recall if g_m is the spectral density of an AR(m) process, then for $n >> m$, $R_n(g_m)^{-1}$ will be bandlimited with most of its rows a shift of the other (thus with the exception of the first m and last m rows it is close to circulant).

3. We approximate $R_n(g_m)^{-1}$ with a circulant matrix, showing that

$$\underline{X}_n' \left[R_n(g_m)^{-1} - C_n(g_m^{-1})\right]\underline{X}_n \to 0,$$

   where $C_n(g_m^{-2})$ is the corresponding circulant matrix (where for $0 < |i-j| \leq m$ and either i or j is greater than m, $(C_n(g^{-1}))_{ij} = 2\sum_{k=|i-j|}^{m} \phi_{m,k}\phi_{m,k-|i-j|+1} - \phi_{m,|i-j|}$) with the eigenvalues $\{g_m(\omega_k)^{-1}\}_{k=1}^n$.

4. These steps show that

$$\underline{X}_n' \left[R_n(f_{\boldsymbol{\theta}})^{-1} - U_n(g_m^{-1})\right]\underline{X}_n \to 0$$

   as $m \to \infty$ as $n \to \infty$, which gives the result.

□

Remark 11.7.1 (A heuristic derivation) *We give a heuristic proof. Using the results in Section 10.2 we have see that $R_n(f_{\boldsymbol{\theta}})$ can be approximately written in terms of the eigenvalue and eigenvectors of the circulant matrix associated with $R_n(f_{\boldsymbol{\theta}})$, that is*

$$R_n(f_{\boldsymbol{\theta}}) \approx F_n \Delta(f_{\boldsymbol{\theta}})\bar{F}_n \quad \textit{thus} \quad R_n(f_{\boldsymbol{\theta}})^{-1} \approx \bar{F}_n \Delta(f_{\boldsymbol{\theta}})^{-1} F_n, \tag{11.33}$$

where $\Delta(f_{\boldsymbol{\theta}}) = diag(f_{\boldsymbol{\theta}}^{(n)}(\omega_1), \ldots, f_{\boldsymbol{\theta}}^{(n)}(\omega_n))$, $f_{\boldsymbol{\theta}}^{(n)}(\omega) = \sum_{j=-(n-1)}^{(n-1)} c_{\boldsymbol{\theta}}(k)\exp(ik\omega) \to f_{\boldsymbol{\theta}}(\omega)$ and

$\omega_k = 2\pi k/n$. *Basic calculations give*

$$\mathbf{X}_n \bar{F}_n = (J_n(\omega_1), \ldots, J_n(\omega_n)). \tag{11.34}$$

Substituting (11.34) and (11.33) into (11.35) yields

$$\frac{1}{n} L_n(\boldsymbol{\theta}) \approx -\frac{1}{n} \sum_{k=1}^{n} \left(\log f_{\boldsymbol{\theta}}(\omega_k) + \frac{|J_n(\omega_k)|^2}{f_{\boldsymbol{\theta}}(\omega_k)} \right) = \frac{1}{n} L^w(\boldsymbol{\theta}). \tag{11.35}$$

Hence using the approximation in (11.33) leads to a heuristic equivalence between the Whittle and Gaussian likelihood.

Chapter 12

Multivariate time series

Objectives

-

12.1 Background

12.1.1 Preliminaries 1: Sequences and functions

Suppose the sequence $a = (\ldots, a_{-2}, a_{-1}, a_0, a_1, \ldots)$ satisfies the property that $\|a\|_2^2 = \sum_{j\in\mathbb{Z}} |a_j|^2 < \infty$. Using a we define the function

$$A(\omega) = \sum_{j\in\mathbb{Z}} a_j \exp(ij\omega).$$

$A(\omega)$ is defined on $[0, 2\pi)$ and has wrapping. In the sense $A(0) = A(2\pi)$. Further $\overline{A(\omega)} = A(-\omega) = A(2\pi - \omega)$. We can extract a_r from $A(\omega)$ using the inverse Fourier transform:

$$a_r = \frac{1}{2\pi}\int_0^{2\pi} A(\omega)e^{-ir\omega}d\omega.$$

To understand why use (12.1) below.

12.1.2 Preliminaries 2: Convolution

We state some well known results on the convolution of sequences, which will be very useful when analysing multivariate time series.

Let ℓ_2 denote the space of square summable sequences. This means if the sequence $a = (\ldots, a_{-2}, a_{-1}, a_0, a_1, \ldots)$ is such that $\|a\|_2^2 = \sum_{j\in\mathbb{Z}} |a_j|^2 < \infty$, then $a \in \ell_2$. Examples of sequences in ℓ_2 is the short memory autocovariance function which is absolutely summable ($\sum_{r\in\mathbb{Z}} |c(r)| < \infty$)

Suppose $a, b \in \ell_2$, and define the convolution

$$\sum_{j\in\mathbb{Z}} a_j b_{j-k}.$$

The convolution can be represented in the Fourier domain. Define

$$A(\lambda) = \sum_{j\in\mathbb{Z}} a_j \exp(-ij\lambda) \quad \text{and} \quad B(\lambda) = \sum_{j\in\mathbb{Z}} b_j \exp(-ij\lambda).$$

Then

$$\sum_{j\in\mathbb{Z}} a_j b_{j-k} = \frac{1}{2\pi}\int_0^{2\pi} \overline{A(\lambda)} B(\lambda) \exp(-ik\lambda) d\lambda.$$

Proof of identity We use the property

$$\frac{1}{2\pi}\int_0^{2\pi} \exp(ij\lambda)d\lambda = \begin{cases} 1 & j = 0 \\ 0 & j \in \mathbb{Z}/\{0\} \end{cases}. \tag{12.1}$$

Expanding the integral gives

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi} \overline{A(\lambda)} B(\lambda) \exp(-ik\lambda)d\lambda &= \sum_{j_1\in\mathbb{Z}}\sum_{j_2\in\mathbb{Z}} a_{j_1} b_{j_2} \underbrace{\frac{1}{2\pi}\int_0^{2\pi} e^{-i(-j_1+j_2+k)\lambda} d\lambda}_{=0 \text{ unless } j_2 = j_1 - k} \\ &= \sum_{j\in\mathbb{Z}} a_j b_{j-k}, \end{aligned}$$

this proves the identity.

12.1.3 Preliminaries 3: Spectral representations and mean squared errors

Univariate times series

We state some useful identities for mean squared errors, first in the univariate case, then in the multivariate case. First we recall the spectral representation of a second order stationary time series (described in Section 10.3.1). Suppose that $\{X_t\}$ is a second order stationary time series. Then it has the representation

$$X_t = \frac{1}{2\pi}\int_0^{2\pi} \exp(it\omega)dZ(\omega),$$

where $\{Z(\omega); \omega \in [0, 2\pi]\}$ is a complex random function that satisfies the orthogonal increment property. We will use the follow properties. If A and B are zero mean, complex random variables, then $\text{cov}[A, B] = \text{E}[A\overline{B}]$ and $\text{var}[A] = \text{cov}(A, A) = \text{E}[A\overline{A}] = \text{E}|A|^2$. Using this we have

- If (ω_1, ω_2) and (ω_3, ω_4) are non-intersection intervals, then $\text{cov}[Z(\omega_2)-Z(\omega_1), Z(\omega_4)-Z(\omega_3)] = 0$.
- Thus $\text{var}[Z(\omega)] = \text{E}[|Z(\omega)|^2] = F(\omega)$, where F is the spectral distribution function (positive non-decreasing function). If $\{X_t\}$ is purely non-deterministic and its autocovariance belongs to ℓ_2, then $F' = f(\omega)$, where $f(\omega) = (2\pi)^{-1}\sum_{r\in\mathbb{Z}} c(r)e^{ir\omega}$.
- Using the above $\text{E}[|dZ(\omega)|^2] = \text{E}[dZ(\omega)\overline{dZ(\omega)}] = dF(\omega) = f(\omega)d\omega$ and $\text{E}[dZ(\omega_1)\overline{dZ(\omega_2)}] = 0$ if $\omega_1 \neq \omega_2$.

Mean squared errors Our aim is to rewrite the mean squared error

$$\text{E}\left(X_t - \sum_{j\neq 0} a_j X_{t-j}\right)^2$$

using the spectral representation. To do so, we replace X_{t-j} in the above with

$$X_{t-j} = \frac{1}{2\pi}\int_0^{2\pi} \exp(i(t-j)\omega)dZ(\omega).$$

We will use the fact that if A is a real random variables then $\text{E}(A) = \text{E}(A\overline{A})$, which will simplify

the calculations. This gives

$$\begin{aligned}
\mathrm{E}\left(X_t - \sum_{j\neq 0} a_j X_{t-j}\right)^2 &= \mathrm{E}\left|\frac{1}{2\pi}\int_0^{2\pi} \exp(it\omega)dZ(\omega) - \sum_{j\neq 0} a_j \frac{1}{2\pi}\int_0^{2\pi}\exp(i(t-j)\omega)dZ(\omega)\right|^2 \\
&= \mathrm{E}\left|\frac{1}{2\pi}\int_0^{2\pi} \exp(it\omega)dZ(\omega) - \sum_{j\neq 0} a_j \frac{1}{2\pi}\int_0^{2\pi}\exp(i(t-j)\omega)dZ(\omega)\right|^2 \\
&= \mathrm{E}\left|\frac{1}{2\pi}\int_0^{2\pi} e^{it\omega}\left(1 - \sum_{j\neq 0} a_j e^{-ij\omega}\right)dZ(\omega)\right|^2 \\
&= \mathrm{E}\left(\frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi} A(\omega_1)\overline{A(\omega_2)}dZ(\omega_1)\overline{dZ(\omega_2)}\right)
\end{aligned}$$

where we set $A(\omega) = 1 - \sum_{j\neq 0} a_j e^{-ij\omega}$. Now we use the property that $\mathrm{E}[|dZ(\omega)|^2] = \mathrm{E}[dZ(\omega)\overline{dZ(\omega)}] == dF(\omega) = f(\omega)d\omega$ and $\mathrm{E}[dZ(\omega_1)\overline{dZ(\omega_2)}] = 0$ if $\omega_1 \neq \omega_2$ this gives

$$\begin{aligned}
\mathrm{E}\left(X_t - \sum_{j\neq 0} a_j X_{t-j}\right)^2 &= \frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi} A(\omega_1)\overline{A(\omega_2)}\mathrm{E}\left(dZ(\omega_1)\overline{dZ(\omega_2)}\right) \\
&= \frac{1}{(2\pi)}\int_0^{2\pi} |A(\omega)|^2 \mathrm{E}|dZ(\omega)|^2 = \frac{1}{(2\pi)}\int_0^{2\pi} |A(\omega)|^2 f(\omega)d(\omega).
\end{aligned}$$

Observe that $|A(\omega)|^2 \geq 0$ for all ω.

Multivariate times series

Suppose $\underline{U}_t' = (X_t, \underline{Y}_t')$ (column vector) is a d-dimensional, zero mean, second order stationary multivariate time series. Second order stationarity in the multivariate set-up is similar to second order stationarity for univariate time series. But there are some important differences. A multivariate time series is second order stationary if for all $t \in \mathbb{Z}$ and $h \in \mathbb{Z}$ we have

$$\mathrm{cov}\left(\underline{U}_t, \underline{U}_{t+h}\right) = C(h), \tag{12.2}$$

where $C(h)$ is a $d \times d$ matrix. However, unlike univariate time series, $C(h) = C(-h)$ does not necessarily hold. Similar to univariate time series, the spectral density matrix is defined as

$$\Sigma(\omega) = \sum_{h\in\mathbb{Z}} C(h)\exp(ih\omega).$$

The matrix $\Sigma(\omega)$ looks like

$$\Sigma(\omega) = \begin{pmatrix} f_{11}(\omega) & f_{12}(\omega) & \dots & f_{1d}(\omega) \\ f_{21}(\omega) & f_{22}(\omega) & \dots & f_{2d}(\omega) \\ \dots & \dots & \ddots & \dots \\ f_{d1}(\omega) & f_{d2}(\omega) & \dots & f_{dd}(\omega) \end{pmatrix}.$$

The diagonal of $\Sigma(\omega)$ is simply the regular univariate spectral density, i.e. $f_{aa}(\omega)$ is the spectral density of the stationary time series $\{X_t^{(a)}\}_t$. The off-diagonal gives cross dependence information. For example by using (12.2) we can see that

$$f_{ab}(\omega) = \sum_{h\in\mathbb{Z}} c_{ab}(h)\exp(ih\omega),$$

where $c_{ab}(h) = \mathrm{cov}(X_t^{(a)}, X_{t+h}^{(b)})$. Thus $f_{ab}(\omega)$ is a frequency measure of cross dependence between two time series. With a litte thought one can see that

$$f_{ba}(\omega) = \overline{f_{ab}(\omega)} = f_{ab}(2\pi - \omega).$$

In other words, $\Sigma(\omega) = \Sigma(\omega)^*$ (where A^* denotes the conjugate transpose of A).

The vectors time series $\{\underline{U}_t\}$ has the representation

$$\underline{U}_t = \frac{1}{2\pi}\int_0^{2\pi} \exp(it\omega)d\underline{Z}_U(\omega), \tag{12.3}$$

where $\underline{Z}_U(\omega)$ is an orthogonal increment process, which has similar properties to the univariate orthogonal increment process. Under summability conditions on the autocovariance matrix we have for $\omega_1 \neq \omega_2$

$$\mathrm{E}\left(d\underline{Z}_U(\omega_1)\overline{d\underline{Z}_U(\omega_2)}\right) = 0$$

and

$$\mathrm{E}\left(d\underline{Z}_U(\omega)\overline{d\underline{Z}_U(\omega)}'\right) = \mathrm{E}\left(d\underline{Z}_U(\omega)d\underline{Z}_U(\omega)^*\right) = \Sigma(\omega)d\omega,$$

where $\underline{V}^*$ denotes the conjugation and transpose of the vector $\underline{V}$.

We write the spectral matrix as the block matrix below

$$\Sigma(\omega) = \begin{pmatrix} f_{XX}(\omega) & f_{X,Y}(\omega) \\ f_{Y,X}(\omega) & f_{YY}(\omega) \end{pmatrix}.$$

Spectral representations of linear sums Using (12.3) we have

$$\begin{aligned} \underline{V}_t = \sum_{j\in\mathbb{Z}} B_j \underline{Y}_{t-j} &= \sum_{j\in\mathbb{Z}} B_j \frac{1}{2\pi}\int_0^{2\pi} e^{i(t-j)\omega} d\underline{Z}_Y(\omega) \\ &= \frac{1}{2\pi}\int_0^{2\pi} e^{it\omega}\left[\sum_{j\in\mathbb{Z}} B_j e^{-ij\omega}\right] d\underline{Z}_Y(\omega) \\ &= \frac{1}{2\pi}\int_0^{2\pi} e^{it\omega} B(\omega) d\underline{Z}_Y(\omega). \end{aligned}$$

Using the same arguments as in the univariate case we have

$$\mathrm{cov}[\underline{V}_t, \underline{V}_{t+h}] = \mathrm{E}[\underline{V}_t \underline{V}_{t+h}'] = \frac{1}{2\pi}\int_0^{2\pi} e^{-ih\omega} B(\omega) f_{YY}(\omega) B(\omega)^* d\omega.$$

Therefore, by the uniqueness of Fourier transforms the spectral density of $\{\underline{V}_t\}$ is $B(\omega)f_{YY}(\omega)B(\omega)^*$.

In order to obtain best linear predictors in a multivariate time series we note that by using the above we can write $X_t - \sum_{j\in\mathbb{Z}} A_j'\underline{Y}_{t-j}$

$$X_t - \sum_{j\in\mathbb{Z}} A_j'\underline{Y}_{t-j} = \frac{1}{2\pi}\int_0^{2\pi} e^{it\omega}\left[dZ_X(\omega) - A(\omega)' d\underline{Z}_Y(\omega)\right], \tag{12.4}$$

where $A(\omega) = \sum_{j\in\mathbb{Z}} A_j e^{-ij\omega}$. This will be our focus below.

Mean squared errors Our aim is to rewrite the mean squared error

$$\mathrm{E}\left(X_t - \sum_{j\in\mathbb{Z}} A_j'\underline{Y}_{t-j}\right)^2$$

using the spectral representation. To do this, we substitute (12.4) into the above to give

$$\begin{aligned}
&\mathrm{E}\left(X_t - \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j}\right)^2 = \mathrm{E}\left|\frac{1}{2\pi}\int_0^{2\pi} e^{it\omega}\left[dZ_X(\omega) - A(\omega)' d\underline{Z}_Y(\omega)\right]\right|^2 \\
= &\frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi} e^{it(\omega_1-\omega_2)}\left[dZ_X(\omega_1) - A(\omega_1)' d\underline{Z}_Y(\omega_1)\right]\left[d\overline{Z_X(\omega_2)} - A(\omega_2)^* d\overline{\underline{Z}_Y(\omega_2)}\right] \\
= &\frac{1}{(2\pi)}\int_0^{2\pi} \mathrm{E}\left[dZ_X(\omega) - A(\omega)' d\underline{Z}_Y(\omega_1)\right]\left[d\overline{Z_X(\omega)} - A(\omega)^* d\overline{\underline{Z}_Y(\omega)}\right],
\end{aligned}$$

where in the last equality of the above we use that $\mathrm{E}\left(d\underline{Z}_U(\omega_1)\overline{d\underline{Z}_U(\omega_2)}\right) = 0$ if $\omega_1 \neq \omega_2$. We note that $\mathrm{E}\left[dZ_X(\omega) - A(\omega)' d\underline{Z}_Y(\omega_1)\right]\left[d\overline{Z_X(\omega)} - A(\omega)^* d\overline{\underline{Z}_Y(\omega)}\right]$ is positive and can be written as

$$\begin{aligned}
&\mathrm{E}\left[dZ_X(\omega) - A(\omega)' d\underline{Z}_Y(\omega)\right]\left[d\overline{Z_X(\omega)} - A(\omega)^* d\overline{\underline{Z}_Y(\omega)}\right] \\
= &\mathrm{E}\left[|dZ_X(\omega)|^2 - A(\omega)' d\underline{Z}_Y(\omega)\right]\left[d\overline{Z_X(\omega)} - A(\omega)^* d\overline{\underline{Z}_Y(\omega)}\right] \\
= &\left[f_{XX}(\omega) - A(\omega)' f_{Y,X}(\omega) - f_{X,Y}(\omega)\overline{A(\omega)} + A(\omega)' f_{YY}(\omega)\overline{A(\omega)}\right] d\omega.
\end{aligned}$$

Substituting this into the integral gives

$$\begin{aligned}
&\mathrm{E}\left(X_t - \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j}\right)^2 \\
= &\frac{1}{(2\pi)}\int_0^{2\pi}\left[f_{XX}(\omega) - A(\omega)' f_{Y,X}(\omega) - f_{X,Y}(\omega)\overline{A(\omega)} + A(\omega)' f_{YY}(\omega)\overline{A(\omega)}\right] d\omega.
\end{aligned}$$

By using a similar argument we have that

$$\begin{aligned}
&\mathrm{cov}\left(X_t - \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j}, X_\tau - \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{\tau-j}\right) \\
= &\frac{1}{(2\pi)}\int_0^{2\pi} \exp(i(t-\tau)\omega)\left[f_{XX}(\omega) - A(\omega)' f_{Y,X}(\omega) - f_{X,Y}(\omega)\overline{A(\omega)} + A(\omega)' f_{YY}(\omega)\overline{A(\omega)}\right] d\omega.
\end{aligned}$$

Thus by the uniqueness of Fourier transforms

$$\left[f_{XX}(\omega) - A(\omega)' f_{Y,X}(\omega) - f_{X,Y}(\omega)\overline{A(\omega)} + A(\omega)' f_{YY}(\omega)\overline{A(\omega)}\right]$$

is the spectral density of the transformed time series $\{X_t - \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j}\}_t$.

12.2 Multivariate time series regression

Suppose that $\{X_t^{(a)}\}$ is a univariate (for simplicity) time series and $\{\underline{Y}_t\}$ a multivariate time series, where $\{X_t^{(a)}, \underline{Y}_t\}_{t\in\mathbb{Z}}$ is jointly second order stationary. Let $\Sigma(\cdot)$ denote the spectral density matrix

$$\Sigma(\omega) = \begin{pmatrix} f_{XX}(\omega) & f_{X,Y}(\omega) \\ f_{Y,X}(\omega) & f_{YY}(\omega) \end{pmatrix}$$

corresponding to $\{X_t^{(a)}, \underline{Y}_t\}_{t\in\mathbb{Z}}$. Our aim is to project $X_t^{(a)}$ onto the space spanned by $\{\underline{Y}_{t-j}\}_{t-j}$. This is not a model but simply a projection. Since $\{X_t^{(a)}, \underline{Y}_t\}$ is second order stationary the regressions will be invariant to shift. That is

$$X_t^{(a)} = \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j} + \varepsilon_t^{(a)}, \tag{12.5}$$

where $\text{cov}[\varepsilon_t^{(a)}, \underline{Y}_{t-j}] = 0$ and the d-dimension vectors A_j do not depend on t. Let

$$A(\omega) = \sum_{j\in\mathbb{Z}} A_j' \exp(-ij\omega).$$

Wiener showed that the $A(\omega)$ which gave the equation (12.5) could easily be solved in the frequency domain with

$$A(\omega) = f_{YY}(\omega)^{-1} f_{Y,X}(\omega), \tag{12.6}$$

and by inverting $A(\omega)$ we can obtain the coefficients $\{A_j\}$;

$$A_j = \frac{1}{2\pi}\int_0^{2\pi} A(\omega)\exp(-ij\omega)d\omega.$$

Using this identity, one can easily obtain an estimator of $A(\omega)$ by estimating $f_{YY}(\cdot)$ and $f_{Y,X}(\cdot)$ and using these estimators to estimate $A(\omega)$. We prove (12.6) in Section **??**. We start with some implications and applications of the regression (12.5).

12.2.1 Conditional independence

In this section we discuss some of the notable features of projecting $X_t^{(a)}$ onto $\{\underline{Y}_t\}$ where $\underline{Y}_t' = (X_t^{(1)}, \ldots, X_t^{(d)})$. We recall that

$$\begin{aligned} X_t^{(a)} &= \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j} + \varepsilon_t^{(a)} \\ &= \sum_{j\in\mathbb{Z}} \sum_{\ell=1}^{d} A_{j,\ell} X_{t-j}^{(\ell)} + \varepsilon_t^{(a)} \end{aligned}$$

Remark 12.2.1 (Special case) *Suppose $A_{j,s} = 0$ for all j. This means the contribution of the time series $\{X_t^{(s)}\}_t$ in "predicting" $X_t^{(a)}$ is zero. In other words, after projecting on the remaining time series, the vector $\underline{Y}_{t,-s}' = (X_t^{(1)}, \ldots, X_t^{(s-1)}, X_t^{(s+1)}, \ldots, X_t^{(d)})$, $\{X_t^{(s)}\}_t$ is conditionally uncorrelated (or independent if Gaussian) with $\{X_t^{(a)}\}$.*

The regression is one-sided. But we know in the multivariate set-up the regression coefficients are related to the partial correlation though the expression identity

$$\beta_{ij} = \rho_{ij} \sqrt{\frac{\mathrm{var}(\varepsilon_i)}{\mathrm{var}(\varepsilon_j)}}. \tag{12.7}$$

We now show a similar expression holds in the context of partial correlation. However, we first need to define the notion of partial correlation and coherency of a time series.

12.2.2 Partial correlation and coherency between time series

We now define the notion of partial correlation between time series and relate it to the regression coefficients in time series, mentioned above.

Typically this is defined in terms of elements in one large vector time series, but to simplify notation suppose $\{(X_t^{(a)}, X_t^{(b)}, \underline{Y}_t')\}_t$ is a $(d+2)$-dimension second order stationary time series. Let

$$\Sigma(\omega) = \begin{pmatrix} f_{aa}(\omega) & f_{ab}(\omega) & f_{a,Y}(\omega) \\ f_{ba}(\omega) & f_{bb}(\omega) & f_{b,Y}(\omega) \\ f_{Y,a}(\omega) & f_{Y,b}(\omega) & f_{YY}(\omega) \end{pmatrix}$$

denote the spectral density matrix. Let

$$\begin{aligned}\varepsilon_{t,Y}^{(a)} &= X_t^{(a)} - \sum_{j=-\infty}^{\infty} (A_j^{(a)})' \underline{Y}_{t-j} \\ \varepsilon_{t,Y}^{(b)} &= X_t^{(b)} - \sum_{j=-\infty}^{\infty} (A_j^{(b)})' \underline{Y}_{t-j}\end{aligned}$$

where $\{A_j^{(a)}\}_j$ and $\{A_j^{(b)}\}_j$ are the coefficients which minimise the MSE and by using (12.6)

$$A^{(a)}(\omega) = f_{YY}(\omega)^{-1} f_{Y,a}(\omega) \text{ and } A^{(b)}(\omega) = f_{YY}(\omega)^{-1} f_{Y,b}(\omega) \tag{12.8}$$

with $A^{(a)}(\omega) = \sum_{j\in\mathbb{Z}} A_j^{(a)} \exp[(ij\omega)$ and $A^{(b)}(\omega) = \sum_{j\in\mathbb{Z}} A_j^{(b)} \exp[(ij\omega)$. We define the partial covariance between $\{\varepsilon_{t,Y}^{(a)}\}_t$ and $\{\varepsilon_{t,Y}^{(b)}\}_t$ as

$$\text{cov}\left[\varepsilon_{t,Y}^{(a)}, \varepsilon_{\tau,Y}^{(a)}\right] = c_{a,b|Y}(t-\tau).$$

Observe that because the original time series is second order stationary, $\varepsilon_{t,Y}^{(a)}$ are also second order stationary (as it is a linear combination of a second order stationary time series time series).

As in the case the regression coefficient is zero for all lags (as discussed in the previous section). If $c_{a,b|Y}(h) = 0$ for all h, then $\{X_t^{(a)}\}$ and $\{X_t^{(b)}\}$ are conditonally independent (when conditioned on $\{\underline{Y}\}_t$).

12.2.3 Cross spectral density of $\{\varepsilon_{t,Y}^{(a)}, \varepsilon_{t,Y}^{(a)}\}$: The spectral partial coherency function

We now evaluate the cross spectral density function of $\{(\varepsilon_{t,Y}^{(a)}, \varepsilon_{t,Y}^{(a)})\}_t$ which is the Fourier transform of $\{c_{a,b|Y}(h)\}_h$.

Remark 12.2.2 (What to look out for) *If $c_{a,b|Y}(h) = 0$ for all h then its spectral density will be zero too.*

To calculate this we use the representations in Section 12.1.3. In particular equation (12.4),

with

$$\begin{aligned} X_t^{(a)} - \sum_{j\in\mathbb{Z}} (A_j^{(a)})' \underline{Y}_{t-j} &= \frac{1}{2\pi}\int_0^{2\pi} e^{it\omega}\left[dZ_a(\omega) - A^{(a)}(\omega)' d\underline{Z}_Y(\omega)\right] \\ X_t^{(b)} - \sum_{j\in\mathbb{Z}} (A_j^{(b)})' \underline{Y}_{t-j} &= \frac{1}{2\pi}\int_0^{2\pi} e^{it\omega}\left[dZ_b(\omega) - A^{(b)}(\omega)' d\underline{Z}_Y(\omega)\right], \end{aligned}$$

where $\{Z_a(\omega), Z_b(\omega), \underline{Z}_b(\omega)\}$ is an orthogonal increment vector process. Thus by using the result in Section 12.1.3 we have

$$\begin{aligned} c_{a,b|Y}(t-\tau) &= \frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi} e^{it\omega_1 - i\tau\omega_2}\mathrm{E}\left(\left[dZ_b(\omega_1) - A^{(b)}(\omega_1)' d\underline{Z}_Y(\omega_1)\right]\left[dZ_b(\omega_2) - A^{(b)}(\omega_2)' d\underline{Z}_Y(\omega_2)\right]\right) \\ &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{2\pi} e^{i(t-\tau)\omega}\mathrm{E}\left(\left[dZ_b(\omega) - A^{(b)}(\omega)' d\underline{Z}_Y(\omega)\right]\overline{\left[dZ_b(\omega) - A^{(b)}(\omega)' d\underline{Z}_Y(\omega)\right]}\right). \end{aligned}$$

Now we focus on evaluating

$$\begin{aligned} & \mathrm{E}\left(\left[dZ_b(\omega) - A^{(b)}(\omega)' d\underline{Z}_Y(\omega)\right]\overline{\left[dZ_a(\omega) - A^{(a)}(\omega)' d\underline{Z}_Y(\omega)\right]}\right) \\ = \; & \mathrm{E}[dZ_a(\omega)\overline{dZ_b(\omega)}] - \mathrm{E}[dZ_a(\omega)\overline{d\underline{Z}_Y(\omega)}']A^{(a)}(\omega)^* - A^{(b)}(\omega)'\mathrm{E}[d\underline{Z}_Y(\omega)\overline{dZ_a(\omega)}] + \\ & A^{(b)}(\omega)'\mathrm{E}[d\underline{Z}_Y(\omega)\overline{d\underline{Z}_Y(\omega)}']A^{(a)}(\omega)^*. \end{aligned}$$

Substituting (12.8) into the above gives

$$\begin{aligned} & \mathrm{E}\left(\left[dZ_b(\omega) - A^{(b)}(\omega)' d\underline{Z}_Y(\omega)\right]\overline{\left[dZ_a(\omega) - A^{(a)}(\omega)' d\underline{Z}_Y(\omega)\right]}\right) \\ = \; & \left[f_{ab}(\omega) - f_{a,Y}(\omega) f_Y(\omega)^{-1} f_{Y,b}(\omega)\right] d\omega. \end{aligned}$$

Altogether this gives the decomposition

$$c_{a,b|Y}(h) = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{2\pi} e^{-ih\omega}\left[f_{ab}(\omega) - f_{a,Y}(\omega) f_Y(\omega)^{-1} f_{Y,b}(\omega)\right] d\omega.$$

Thus by the uniqueness of the spectral density function, the spectral density of $\{c_{a,b|Y}(h)\}$ is

$$f_{a,b|Y}(\omega) = \sum_{h\in\mathbb{Z}} c_{a,b|Y}(h)\exp(ih\omega) = f_{ab}(\omega) - f_{a,Y}(\omega) f_Y(\omega)^{-1} f_{Y,b}(\omega).$$

The standardized version of $f_{a,b|Y}(\omega)$:

$$\rho_{a,b|Y}(\omega) = \frac{f_{ab}(\omega) - f_{a,Y}(\omega) f_Y(\omega)^{-1} f_{Y,b}(\omega)}{\sqrt{f_{a,a|Y}(\omega) f_{b,b|Y}(\omega)}}$$

is often called the spectral partial coherence in multivariate time series.

12.3 Properties of the inverse of the spectral density matrix

We now show that results analogous to those in multivariate analysis, also apply to the inverse spectral density matrix. In particular, the linear regression $A(\omega)$ and partial coherency $f_{a,b|Y}$ are found within the inverse spectral density matrix. To do this we write everything in terms of the multivariate time series $\{\underline{Y}_t\}_t$.

The important results that we will use is that if Σ is a positive definite matrix and $\Gamma = \Sigma^{-1}$ (it can be complex). Then

- The diagonal

$$\Gamma_{aa} = \frac{1}{\Sigma_{aa} - \Sigma_{a,-a}\Sigma_{-i,-i}\Sigma'_{a,-a}},$$

  where $\Sigma_{a,-a}$ denotes the ath row of Σ with the ath column removed.

- The off-diagonal For $a \neq b$ we have

$$\Gamma_{ab} = -\beta_{a,b}\Gamma_{aa}$$

  where

$$\boldsymbol{\beta}_a = (\beta_{a,1}, \ldots, \beta_{a,d})' = [\Sigma_{-a,-a}]^{-1}\Sigma_{-a,a}. \tag{12.9}$$

- The formula for partial correlation. The partial covariance is

$$\gamma_{a,b} = \Sigma_{a,b} - \Sigma'_{a,-(a,b)}[\Sigma_{-(a,b)}]^{-1}\Sigma_{a,-(a,b)}. \tag{12.10}$$

Thus the partial correlation is

$$\rho_{a,b} = \frac{\gamma_{a,b}}{\sqrt{\gamma_{aa}\gamma_{bb}}}.$$

The regression coefficients and partial correlations are linked by $\rho_{a,b} = \beta_{a,b}\sqrt{\frac{\Gamma_{aa}}{\Gamma_{bb}}}$. Using this we can represent the partial correlation as

$$\rho_{a,b} = -\frac{\Gamma_{a,b}}{\sqrt{\Gamma_{aa}\Gamma_{bb}}}.$$

We will compare the above formulas with the time series regression coefficients (the Fourier transform) and spectral coherency. We will show that these quantities are hidden in the inverse of the spectral density matrix. Let us suppose that $\{\underline{Y}_t\}_t$ is a second order stationary time series, where $\underline{Y}_t' = (X_t^{(1)}, \ldots, X_t^{(d)})$. The corresponding spectral density matrix is

$$\Sigma(\omega) = \begin{pmatrix} f_{11}(\omega) & f_{12}(\omega) & \ldots & f_{1d}(\omega) \\ f_{21}(\omega) & f_{22}(\omega) & \ldots & f_{2d}(\omega) \\ \ldots & \ldots & \ddots & \ldots \\ f_{d1}(\omega) & f_{d2}(\omega) & \ldots & f_{dd}(\omega) \end{pmatrix}.$$

We let $\Gamma(\omega) = \Sigma(\omega)^{-1}$.

Formulas for time series regression coefficients and spectral coherency:

- The best linear predictor of $X_t^{(a)}$ given $\{\underline{Y}_t^{(-a)}\}_t$ is

$$X_t^{(a)} = \sum_{\ell \in \mathbb{Z}} (A_\ell^{(a)})' \underline{Y}_{t-\ell}^{(-a)} + \varepsilon_t^{(a)}.$$

Let $A^{(a)}(\omega) = \sum_{\ell \in \mathbb{Z}} A_\ell^{(a)} \exp(i\ell\omega)$, then

$$A^{(a)}(\omega) = [\Sigma(\omega)_{-a,-a}]^{-1} \Sigma(\omega)_{-a,a}, \tag{12.11}$$

where $\Sigma(\omega)_{-a,-a}$ is $\Sigma(\omega)$ but with the jth column and row removed.

- Formulas for spectral coherency (Fourier transform of partial covariance) between $\{X_t^{(a)}\}_t$

and $\{X_t^{(b)}\}_t$ given $\{\underline{Y}_t^{-(a,b)}\}_t$ is

$$f_{a,b|-(a,b)}(\omega) = \Sigma_{ab}(\omega) - \Sigma_{a,-(a,b)}(\omega)[\Sigma_{-(a,b),-(a,b)}(\omega)]^{-1}\Sigma_{-(a,b),b}(\omega) \tag{12.12}$$

where $\Sigma_{a,-(a,b)}(\omega)$ is the ath row of $\Sigma(\omega)$ but with the (a,b)th column removed. Using the above we define

$$\rho_{a,b}(\omega) = \frac{f_{ab}(\omega) - f_{a,-(a,b)}(\omega)f_{-(a,b)}(\omega)^{-1}f_{-(a,b),b}(\omega)}{\sqrt{f_{a,a|-(a,b)}(\omega)f_{b,b|-(a,b)}(\omega)}}.$$

Making the comparison Comparing (12.9) with (12.11), we observe that

$$\Gamma(\omega)_{a,a} = -\frac{1}{f_{-a}(\omega)} \quad \text{where } f_{-a}(\omega) = \Sigma(\omega)_{a,a} - \Sigma(\omega)_{a,-a}[\Sigma(\omega)_{-a,-a}]^{-1}\Sigma_{a,-a}(\omega)'.$$

For $a \neq b$

$$\Gamma(\omega)_{a,b} = -[A^{(a)}(\omega)]_b\Gamma(\omega)_{a,a}. \tag{12.13}$$

Comparing $f_{-a}(\omega)$ with Section 12.1.3 it can be seen that $f_{-a}(\omega)$ is the spectral density of the residual time series $\{X_t^{(a)} - \sum_\ell (A_j^{(a)})'\underline{Y}_{t-\ell}^{(-a)}\}$. This links the regression coefficients in time series with those of the precision matrix.

We now turn to partial spectral coherency. Again comparing (12.10) with (12.12) we can see that

$$\rho_{a,b}(\omega) = -\frac{\Gamma(\omega)_{a,b}}{\sqrt{\Gamma(\omega)_{aa}\Gamma(\omega)_{bb}}}. \tag{12.14}$$

In conclusion we have shown that all the formulas which connect the precision matrix to linear regression and partial correlation in multivariate analysis transfer to the precision spectral density matrix in stationary time series. However, the results derived above are based simply on a comparision of formulas. Below we give an intuition why these must hold using the orthogonal increment process $\{\underline{Z}(\omega); \omega \in [0, 2\pi)\}$.

A heuristic understanding in terms of the orthogonal increment process

Let $\underline{Y}_t' = (X_t^{(1)}, \ldots, X_t^{(d)})$, and suppose $\{\underline{Y}_t\}_t$ is a second order stationary time series. Then it has the spectral representation

$$\underline{Y}_t = \frac{1}{2\pi} \int_0^{2\pi} \exp(it\omega) d\underline{Z}_Y(\omega),$$

where $\underline{Z}_Y(\omega)$ is an orthogonal increment process. Very roughly speaking (and totally ignoring Ito Calculus) this means we can treat the increments/segments $\{\underline{\Delta}_k\}_{k=0}^{2\pi/\delta-1}$ where $\underline{\Delta}_k = \underline{Z}((k+1)\delta) - \underline{Z}(k\delta)$ as roughly uncorrelated (but not identically distributed) random vectors. The random vector $\underline{\Delta}_k$ has mean zero and variance that is

$$\mathrm{var}[\underline{\Delta}_k] = \Sigma((k+1)\delta) - \Sigma(k\delta) \approx \delta\Sigma(k\delta).$$

Now we can apply all the results in multivariate analysis to the random vector $\underline{\Delta}_k$ (regression and partial correlation) and obtain the formulas above.

12.4 Proof of equation (12.6)

We recall in equation (12.5) we stated that the best linear predictor of $X_t^{(a)}$ given $\underline{Y}_t$ is

$$X_t^{(a)} = \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j} + \varepsilon_t^{(a)},$$

where

$$A(\omega) = f_{YY}(\omega)^{-1} f_{Y,X}(\omega),$$

$A(\omega) = \sum_{j\in\mathbb{Z}} A_j' \exp(-ij\omega)$. We now prove this result. There are various method but in this section we use the spectral representation which reduces the mean squared error

$$\mathrm{E}\left(X_t - \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j} \right)^2.$$

By using the results in Section 12.1.3 we can write the MSE as

$$\begin{aligned}
\mathrm{E}\left(X_t - \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j}\right)^2 &= \mathrm{E}\left|\frac{1}{2\pi}\int_0^{2\pi} e^{it\omega}\left[dZ_X(\omega) - A(\omega)' d\underline{Z}_Y(\omega)\right]\right|^2 \\
&= \frac{1}{(2\pi)}\int_0^{2\pi} \mathrm{E}\left[dZ_X(\omega) - A(\omega)' d\underline{Z}_Y(\omega_1)\right]\left[d\overline{Z_X(\omega)} - A(\omega)^* d\overline{\underline{Z}_Y(\omega)}\right],
\end{aligned}$$

where

$$\begin{aligned}
&\mathrm{E}\left[dZ_X(\omega) - A(\omega)' d\underline{Z}_Y(\omega)\right]\left[d\overline{Z_X(\omega)} - A(\omega)^* d\overline{\underline{Z}_Y(\omega)}\right] \\
&= \left[f_{XX}(\omega) - A(\omega)' f_{Y,X}(\omega) - f_{X,Y}(\omega)\overline{A(\omega)} + A(\omega)' f_{YY}(\omega)\overline{A(\omega)}\right] d\omega.
\end{aligned}$$

Substituting this into the integral gives

$$\begin{aligned}
&\mathrm{E}\left(X_t - \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j}\right)^2 \\
&= \frac{1}{(2\pi)}\int_0^{2\pi} \left[f_{XX}(\omega) - A(\omega)' f_{Y,X}(\omega) - f_{X,Y}(\omega)\overline{A(\omega)} + A(\omega)' f_{YY}(\omega)\overline{A(\omega)}\right] d\omega. \quad (12.15)
\end{aligned}$$

Studying the above expansion we observe that the term inside the integral

$$L(A;\omega) = \left[f_{XX}(\omega) - A(\omega)' f_{Y,X}(\omega) - f_{X,Y}(\omega)\overline{A(\omega)} + A(\omega)' f_{YY}(\omega)\overline{A(\omega)}\right]$$

is non-negative. Thus we can find the $A(\omega)$ which minimises (12.15) by finding the $A(\omega)$ which minimises $L(A;\omega)$ for each $\omega \in [0,\pi]$ (since $\overline{A(\omega)} = A(2\pi - \omega)$). We first note that the vector $A(\omega)$ is a complex vector, thus we partition it in terms of its real and imaginary parts

$$A(\omega) = \underline{a}(\omega) + i\underline{b}(\omega).$$

Substituting this into $L(A;\omega)$ and differentiating with respect to the entries in $\underline{a}(\omega)$ and $\underline{b}(\omega)$ gives

$$\begin{aligned}
\nabla_{\underline{a}} L(A;\omega) &= -f_{Y,a}(\omega) - f_{a,Y}(\omega)^* + f_{YY}(\omega)A(\omega) + (A(\omega)^* f_{YY}(\omega))^* \\
\nabla_{\underline{b}} L(A;\omega) &= -if_{Y,a}(\omega) + if_{a,Y}(\omega)^* + if_{YY}(\omega)A(\omega) - i(A(\omega)^* f_{YY}(\omega))^*.
\end{aligned}$$

Equating the derivatives to zero and solving for $\underline{a}(\omega)$ and $\underline{b}(\omega)$ gives

$$A(\omega) = f_{YY}(\omega)^{-1} f_{Y,X}(\omega).$$

Thus we have proved the required result.

An alternative proof involves the normal equations (derivatives of the MSE):

$$\mathrm{E}\left[\left(X_t - \sum_{j\in\mathbb{Z}} A_j' \underline{Y}_{t-j}\right)\underline{Y}_{t-\ell}\right]^2.$$

This avoids going immediately going into the frequency domain (we have done above) and thus the need to take complex derivatives. One then replaces the autocovariances in the above with the spectral density function. It has the advantage of avoiding the need to consider the derivatives of real and imaginary parts (or their complex derivatives). We leave the details as a exercise.

Chapter 13

Nonlinear Time Series Models

Prerequisites

- A basic understanding of expectations, conditional expectations and how one can use conditioning to obtain an expectation.

Objectives:

- Use relevant results to show that a model has a stationary, solution.
- Derive moments of these processes.
- Understand the differences between linear and nonlinear time series.

So far we have focused on linear time series, that is time series which have the representation

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}, \tag{13.1}$$

where $\{\varepsilon_t\}$ are iid random variables. Such models are extremely useful, because they are designed to model the autocovariance structure and are straightforward to use for forecasting. These are some of the reasons that they are used widely in several applications. Note that all stationary Gaussian time series have a linear form (of the type given in (13.1)), where the innovations $\{\varepsilon_t\}$ are Gaussian.

A typical realisation from a linear time series, will be quite regular with no suddent bursts or jumps. This is due to the linearity of the system. However, if one looks at financial data, for example, there are sudden bursts in volatility (variation) and extreme values, which calm down

after a while. It is not possible to model such behaviour well with a linear time series. In order to capture 'nonlinear behaviour several nonlinear models have been proposed. The models typically consists of products of random variables which make possible the sudden irratic bursts seen in the data. Over the past 30 years there has been a lot research into nonlinear time series models. Probably one of the first nonlinear models proposed for time series analysis is the bilinear model, this model is used extensively in signal processing and engineering. A popular model for modelling financial data are (G)ARCH-family of models. Other popular models are random autoregressive coefficient models and threshold models, to name but a few (see, for example, Subba Rao (1977), Granger and Andersen (1978), Nicholls and Quinn (1982), Engle (1982), Subba Rao and Gabr (1984), Bollerslev (1986), Terdik (1999), Fan and Yao (2003), Straumann (2005) and Douc et al. (2014)).

Once a model has been defined, the first difficult task is to show that it actually has a solution which is almost surely finite (recall these models have dynamics which start at the $-\infty$, so if they are not well defined they could be 'infinite'), with a stationary solution. Typically, in the nonlinear world, we look for causal solutions. I suspect this is because the mathematics behind existence of non-causal solution makes the problem even more complex.

We state a result that gives sufficient conditions for a stationary, causal solution of a certain class of models. These models include ARCH/GARCH and Bilinear models. We note that the theorem guarantees a solution, but does not give conditions for it's moments. The result is based on Brandt (1986), but under stronger conditions.

Theorem 13.0.1 (Brandt (1986)) *Let us suppose that $\{\boldsymbol{X}_t\}$ is a d-dimensional time series defined by the stochastic recurrence relation*

$$\boldsymbol{X}_t = A_t \boldsymbol{X}_{t-1} + \boldsymbol{B}_t, \tag{13.2}$$

where $\{A_t\}$ and $\{B_t\}$ are iid random matrices and vectors respectively. If $\mathrm{E}\log\|A_t\| < 0$ and $\mathrm{E}\log\|\boldsymbol{B}_t\| < \infty$ (where $\|\cdot\|$ denotes the spectral norm of a matrix), then

$$\boldsymbol{X}_t = \boldsymbol{B}_t + \sum_{k=1}^{\infty}\left(\prod_{i=0}^{k-1} A_{t-i}\right)\boldsymbol{B}_{t-k} \tag{13.3}$$

converges almost surely and is the unique strictly stationary causal solution.

Note: The conditions given above are very strong and Brandt (1986) states the result under

which weaker conditions, we outline the differences here. Firstly, the assumption $\{A_t, B_t\}$ *are iid can be relaxed to their being Ergodic sequences. Secondly, the assumption* $\mathrm{E}\log\|A_t\| < 0$ *can be relaxed to* $\mathrm{E}\log\|A_t\| < \infty$[1] *and that* $\{A_t\}$ *has a negative Lyapunov exponent, where the Lyapunov exponent is defined as* $\lim_{n\to\infty}\frac{1}{n}\|\prod_{j=1}^{n} A_j\| = \gamma$*, with* $\gamma < 0$ *(see Brandt (1986)).*

The conditions given in the above theorem may appear a little cryptic. However, the condition $\mathrm{E}\log|A_t| < 0$ (in the unvariate case) becomes quite clear if you compare the SRE model with the AR(1) model $X_t = \rho X_{t-1} + \varepsilon_t$, where $|\rho| < 1$ (which is the special case of the SRE, where the coefficients is deterministic). We recall that the solution of the AR(1) is $X_t = \sum_{k=1}^{\infty}\rho^j\varepsilon_{t-j}$. The important part in this decomposition is that $|\rho^j|$ decays geometrically fast to zero. Now let us compare this to (13.3), we see that ρ^j plays a similar role to $\prod_{i=0}^{k-1} A_{t-i}$. Given that there are similarities between the AR(1) and SRE, it seems reasonable that for (13.3) to converge, $\prod_{i=0}^{k-1} A_{t-i}$ should converge geometrically too (at least almost surely). However analysis of a product is not straight forward, therefore we take logarithms to turn it into a sum

$$\frac{1}{k}\log\prod_{i=0}^{k-1} A_{t-i} = \frac{1}{k}\sum_{i=0}^{k-1}\log A_{t-i} \overset{\text{a.s.}}{\to} \mathrm{E}[\log A_t] := \gamma,$$

since it is the sum of iid random variables. Thus taking anti-logs

$$\prod_{i=0}^{k-1} A_{t-i} \approx \exp[k\gamma],$$

which only converges to zero if $\gamma < 0$, in other words $\mathrm{E}[\log A_t] < 0$. Thus we see that the condition $\mathrm{E}\log|A_t| < 0$ is quite a logical conditional afterall.

13.0.1 Examples

The AR(1) model

It is straightforward to see that the causal, stationary AR(1) model satisfies the conditions in Theorem 13.0.1. Observe that since

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

has a stationary causal solution when $|\phi| < 1$, then $\mathrm{E}[\log|\phi|] = \log|\phi| < 0$ (since $|\phi < 1|$).

[1] Usually we use the spectral norm, which is defined as the $\sqrt{\lambda_{\max}(A'A)}$

The AR(2) model

Things become a little tricker with the AR(2) case. We recall from Section **??** that the causal AR(2) model

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

can be written as an VAR(1) model

$$\underline{X}_t = A\underline{X}_{t-1} + \underline{\varepsilon}_t = \sum_{j=0}^{\infty} A^j \underline{\varepsilon}_{t-j} \tag{13.4}$$

where

$$\begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}, \tag{13.5}$$

and $\underline{\varepsilon}_t' = (\varepsilon_t, 0)$. For the process to be causal the eigenvalues of the matrix A should be less than one. Thus in the above example $\phi_1 = a + b$ and $\phi_2 = -ab$, for some $|a|, |b| < 1$. This implies that the eigenvalues of A will be less than one, however the eigenvalues of $A'A$ may not be less than one. For example consider the AR(2) model

$$X_t = 2 \times 0.2 X_{t-1} - 0.2^2 X_{t-2} + \varepsilon_t$$

which correspond to A with the eigenvalues 0.2 and 0.2.

```
A = matrix(c(2*phi,-phi**2,1,0),byrow =T, ncol = 2)
>eigen(A)
eigen() decomposition
$values
[1] 0.2+0i 0.2-0i
> eigen(A%*%t(A))
eigen() decomposition
$values
[1] 1.160220952 0.001379048
```

From the code above, we see that the spectral radius of A (largest eigenvalue of A) is 0.2, but

$\|A\|_{spec} = 1.16$. However, if we evaluate the spectral norm of A^2, it is less than one;

```
> A2 = A%*%A
> eigen(A2%*%t(A2))
eigen() decomposition
$values
[1] 1.762415e-01 1.452553e-05
```

In this example we see that $\|A^2\|_{spec} = \sqrt{0.176}$. Since we can group the product A^k into the products of A^2, this is what gives the contraction. This will happen for any matrix, A, whose eigenvalues are less than one. For a large enough k, the spectral norm of A^k will be less than one[2]. Therefore the conditions of Theorem 13.0.1 are not satisfied. But the weaker conditions (given below the main conditions of the theorem) is satisfied.

Nonlinear Time series models There are many other (nonlinear) models that have the representation in (13.2). The purpose of this chapter is to introduce and motivate some of these models.

13.1 Data Motivation

13.1.1 Yahoo data from 1996-2014

We consider here the closing share price of the Yahoo daily data downloaded from `https://uk.finance.yahoo.com/q/hp?s=YHOO`. The data starts from from 10th April 1996 to 8th August 2014 (over 4000 observations). A plot is given in Figure 13.1. Typically the logarithm of such data taken, and in order to remove linear and/or stochastic trend the first difference of the logarithm is taken (ie. $X_t = \log S_t - \log S_{t-1}$). The hope is that after taking differences the data has been stationarized (see Example 4.7). However, the data set spans almost 20 years and this assumption is rather precarious and will be investigated later. A plot of the data after taking first differences together with the QQplot is given in Figure 13.2. From the QQplot in Figure 13.2, we observe that log differences $\{X_t\}$ appears to have very thick tails, which may mean that higher order moments of the log returns do not exist (not finite).

In Figure 13.3 we give the autocorrelation (ACF) plots of the log differences, absolute log

[2]This result is due to Gelfand's lemma

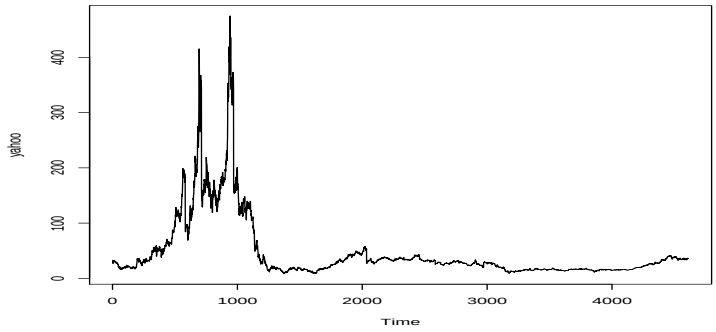


Figure 13.1: Plot of daily closing Yahoo share price 1996-2014

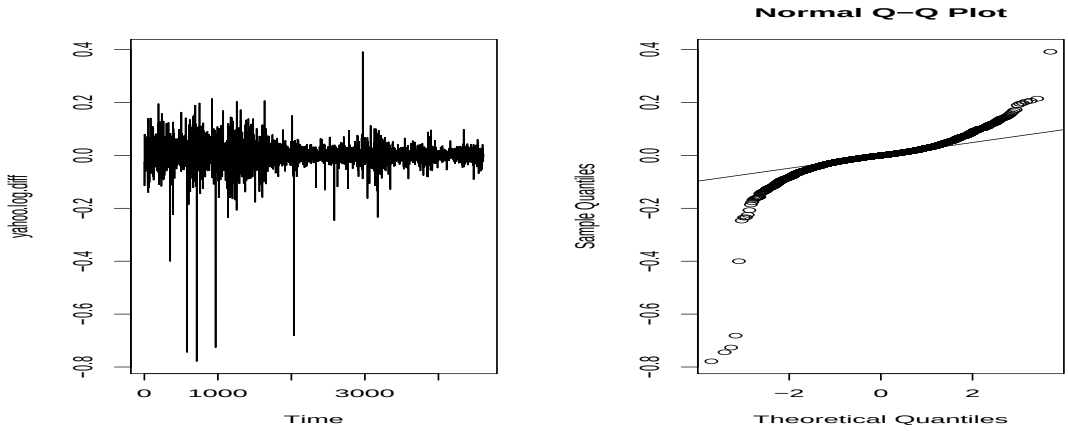


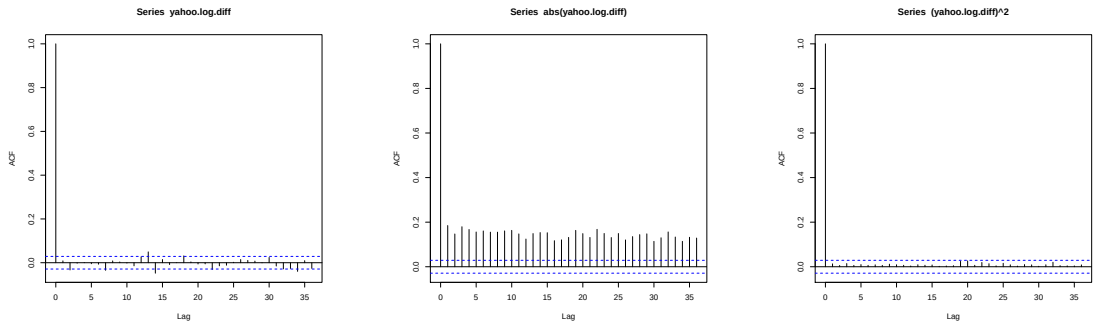
Figure 13.2: Plot of log differences of daily Yahoo share price 1996-2014 and the corresponding QQplot

differences and squares of the log differences. Note that the sample autocorrelation is defined as

$$\widehat{\rho}(k) = \frac{\widehat{c}(k)}{\widehat{c}(0)}, \quad \text{where} \quad \widehat{c}(k) = \frac{1}{T}\sum_{t=1}^{T-|k|}(X_t - \bar{X})(X_{t+k} - \bar{X}). \tag{13.6}$$

The dotted lines are the errors bars (the 95% confidence of the sample correlations constructed under the assumption the observations are independent, see Section 8.2.1). From Figure 13.3a we see that there appears to be no correlation in the data. More precisely, most of the sample correlations are within the errors bars, the few that are outside it could be by chance, as the error bars are constructed pointwise. However, Figure 13.3b the ACF plot of the absolutes gives significant large correlations. In contrast, in Figure 13.3c we give the ACF plot of the squares, where there does not appear to be any significant correlations.

To summarise, $\{X_t\}$ appears to be uncorrelated (white noise). However, once absolutes have



(a) ACF plot of the log differences (b) ACF plot of the absolute of the log differences (c) ACF plot of the square of the log differences

Figure 13.3: ACF plots of the transformed Yahoo data

been taken there does appear to be dependence. This type of behaviour cannot be modelled with a linear model. What is quite interesting is that there does not appear to be any significant correlation in the squares. However, on explanation for this can be found in the QQplot. The data has extremely thick tails which suggest that the forth moment of the process may not exist (the empirical variance of X_t will be extremely large). Since correlation is defined as (13.6) involves division by $\widehat{c}(0)$, which could be extremely large, this would 'hide' the sample covariance.

R code for Yahoo data

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

```
yahoo <- scan("~/yahoo304.96.8.14.txt")
yahoo <- yahoo[c(length(yahoo):1)] # switches the entries to ascending order 1996-2014
yahoo.log.diff <- log(yahoo[-1]) - log(yahoo[-length(yahoo)])
# Takelog differences
par(mfrow=c(1,1))
plot.ts(yahoo)
par(mfrow=c(1,2))
plot.ts(yahoo.log.diff)
qqnorm(yahoo.log.diff)
qqline(yahoo.log.diff)
par(mfrow=c(1,3))
acf(yahoo.log.diff) # ACF plot of log differences
```

```
acf(abs(yahoo.log.diff)) # ACF plot of absolute log differences
acf((yahoo.log.diff)**2) # ACF plot of square of log differences
```

13.1.2 FTSE 100 from January - August 2014

For completeness we discuss a much shorter data set, the daily closing price of the FTSE 100 from 20th January - 8th August, 2014 (141 observations). This data was downloaded from `http://markets.ft.com/research//Tearsheets/PriceHistoryPopup?symbol=FTSE:FSI`.

Exactly the same analysis that was applied to the Yahoo data is applied to the FTSE data and the plots are given in Figure 13.4-13.6.

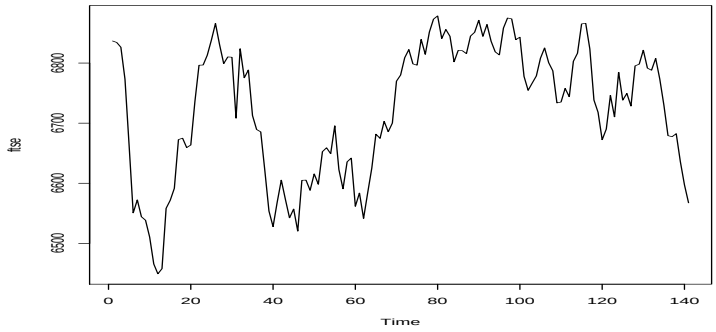


Figure 13.4: Plot of daily closing FTSE price Jan-August, 2014

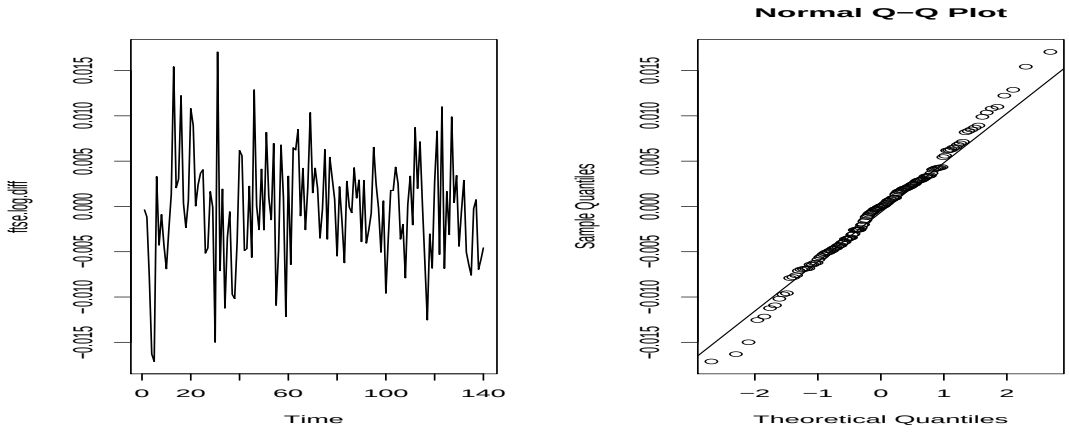
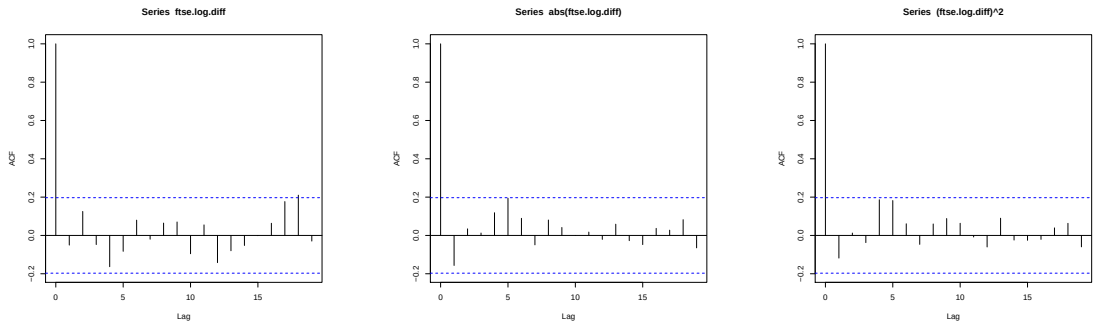


Figure 13.5: Plot of log differences of daily FTSE price Jan-August, 2014 and the corresponding QQplot

We observe that for this (much shorter) data set, the marginal observations do not appear to deviate much from normality (note just because the marginal is Gaussian does not mean the entire



(a) ACF plot of the log differences (b) ACF plot of the absolute of the log differences (c) ACF plot of the square of the log differences

Figure 13.6: ACF plots of the transformed FTSE data

time series is Gaussian). Furthermore, the ACF plot of the log differences, absolutes and squares do not suggest any evidence of correlation. Could it be, that after taking log differences, there is no dependence in the data (the data is a realisation from iid random variables). Or that there is dependence but it lies in a 'higher order structure' or over more sophisticated transformations.

Comparing this to the Yahoo data, may be we 'see' dependence in the Yahoo data because it is actually nonstationary. The mystery continues (we look into this later). It would be worth while conducting a similar analysis on a similar portion of the Yahoo data.

13.2 The ARCH model

During the early 80s Econometricians were trying to find a suitable model for forecasting stock prices. They were faced with data similar to the log differences of the Yahoo data in Figure 13.2. As Figure 13.3a demonstrates, there does not appear to be any linear dependence in the data, which makes the best linear predictor quite useless for forecasting. Instead, they tried to predict the variance of future prices given the past, $\mathrm{var}[X_{t+1}|X_t, X_{t-1}, \ldots]$. This called for a model that has a zero autocorrelation function, but models the conditional variance.

To address this need, Engle (1982) proposed the autoregressive conditionally heteroskadastic (ARCH) model (note that Rob Engle, together with Clive Granger, in 2004, received the Noble prize for Economics for Cointegration). He proposed the ARCH(p) which satisfies the representation

$$X_t = \sigma_t Z_t \qquad \sigma_t^2 = a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2,$$

where Z_t are iid random variables where $\mathrm{E}(Z_t) = 0$ and $\mathrm{var}(Z_t) = 1$, $a_0 > 0$ and for $1 \leq j \leq p$ $a_j \geq 0$.

Before, worrying about whether a solution of such a model exists, let us consider the reasons behind why this model was first proposed.

13.2.1 Features of an ARCH

Let us suppose that a causal, stationary solution of the ARCH model exists (X_t is a function of $Z_t, Z_{t-1}, Z_{t-1}, \ldots$) and all the necessary moments exist. Then we obtain the following.

(i) The first moment:

$$\begin{aligned} \mathrm{E}[X_t] &= \mathrm{E}[Z_t\sigma_t] = \mathrm{E}[\mathrm{E}(Z_t\sigma_t|X_{t-1}, X_{t-2}, \ldots)] = \underbrace{\mathrm{E}[\sigma_t\mathrm{E}(Z_t|X_{t-1}, X_{t-2}, \ldots)]}_{\sigma_t \text{ function of } X_{t-1},\ldots,X_{t-p}} \\ &= \mathrm{E}[\sigma_t \underbrace{\mathrm{E}(Z_t)}_{\text{by causality}}] = \mathrm{E}[0 \cdot \sigma_t] = 0. \end{aligned}$$

Thus the ARCH process has a zero mean.

(ii) The conditional variance:

$$\begin{aligned} \mathrm{var}(X_t|X_{t-1}, X_{t-2}, \ldots, X_{t-p}) &= \mathrm{E}(X_t^2|X_{t-1}, X_{t-2}, \ldots, X_{t-p}) \\ &= \mathrm{E}(Z_t^2\sigma_t^2|X_{t-1}, X_{t-2}, \ldots, X_{t-p}) = \sigma_t^2\mathrm{E}[Z_t^2] = \sigma_t^2. \end{aligned}$$

Thus the conditional variance is $\sigma_t^2 = a_0 + \sum_{j=1}^p a_j X_{t-j}^2$ (a weighted sum of the squared past).

(iii) The autocovariance function:

Without loss of generality assume $k > 0$

$$\begin{aligned} \mathrm{cov}[X_t, X_{t+k}] &= \mathrm{E}[X_tX_{t+k}] = \mathrm{E}[X_t\mathrm{E}(X_{t+k}|X_{t+k-1}, \ldots, X_t)] \\ &= \mathrm{E}[X_t\sigma_{t+k}\mathrm{E}(Z_{t+k}|X_{t+k-1}, \ldots, X_t)] = \mathrm{E}[X_t\sigma_{t+k}\mathrm{E}(Z_{t+k})] = \mathrm{E}[X_t\sigma_{t+k} \cdot 0] = 0. \end{aligned}$$

The autocorrelation function is zero (it is a white noise process).

(iv) We will show in Section 13.2.2 that $\mathrm{E}[X^{2d}] < \infty$ iff $[\sum_{j=1}^p a_j]\mathrm{E}[Z_t^{2d}]^{1/d} < 1$. It is well known that even for Gaussian innovations $\mathrm{E}[Z_t^{2d}]^{1/d}$ grows with d, therefore if any of the a_j are

non-zero (recall all need to be positive), there will exist a d_0 such that for all $d \geq d_0$ $\mathrm{E}[X_t^d]$ will not be finite. Thus the we see that the ARCH process has thick tails.

Usually we measure the thickness of tails in data using the Kurtosis measure (see wiki).

Points (i-iv) demonstrate that the ARCH model is able to model many of the features seen in the stock price data.

In some sense the ARCH model can be considered as a generalisation of the AR model. That is the squares of ARCH model satisfy

$$X_t^2 = \sigma^2 Z_t^2 = a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2 + (Z_t^2 - 1)\sigma_t^2, \tag{13.7}$$

with characteristic polynomial $\phi(z) = 1 - \sum_{j=1}^{p} a_j z^j$. It can be shown that if $\sum_{j=1}^{p} a_j < 1$, then the roots of the characteristic polynomial $\phi(z)$ lie outside the unit circle (see Exercise 4.2). Moreover, the 'innovations' $\epsilon_t = (Z_t^2 - 1)\sigma_t^2$ are *martingale differences* (see wiki). This can be shown by noting that

$$\mathrm{E}[(Z_t^2 - 1)\sigma_t^2 | X_{t-1}, X_{t-2}, \ldots] = \sigma_t^2 \mathrm{E}(Z_t^2 - 1 | X_{t-1}, X_{t-2}, \ldots) = \sigma_t^2 \underbrace{\mathrm{E}(Z_t^2 - 1)}_{=0} = 0.$$

Thus $\mathrm{cov}(\epsilon_t, \epsilon_s) = 0$ for $s \neq t$. Martingales are a useful asymptotic tool in time series, we demonstrate how they can be used in Chapter 14.

To summarise, in many respects the ARCH(p) model resembles the AR(p) except that the innovations $\{\epsilon_t\}$ are martingale differences and not iid random variables. This means that despite the resemblence, it is not a linear time series.

We show that a unique, stationary causal solution of the ARCH model exists and derive conditions under which the moments exist.

13.2.2 Existence of a strictly stationary solution and second order stationarity of the ARCH

To simplify notation we will consider the ARCH(1) model

$$X_t = \sigma_t Z_t \qquad \sigma_t^2 = a_0 + a_1 X_{t-1}^2. \tag{13.8}$$

It is difficult to directly obtain a solution of X_t, instead we obtain a solution for σ_t^2 (since X_t can immediately be obtained from this). Using that $X_{t-1}^2 = \sigma_{t-1}^2 Z_{t-1}^2$ and substituting this into (13.8) we obtain

$$\sigma_t^2 = a_0 + a_1 X_{t-1}^2 = (a_1 Z_{t-1}^2)\sigma_{t-1}^2 + a_0. \tag{13.9}$$

We observe that (13.9) can be written in the stochastic recurrence relation form given in (13.2) with $A_t = a_1 Z_{t-1}^2$ and $B_t = a_0$. Therefore, by using Theorem 13.0.1, if $\mathrm{E}[\log a_1 Z_{t-1}^2] = \log a_1 + \mathrm{E}[\log Z_{t-1}^2] < 0$, then σ_t^2 has the strictly stationary causal solution

$$\sigma_t^2 = a_0 + a_0 \sum_{k=1}^{\infty} a_1^k \prod_{j=1}^{k} Z_{t-j}^2.$$

The condition for *existence* using Theorem 13.0.1 and (13.9) is

$$\mathrm{E}[\log(a_1 Z_t^2)] = \log a_1 + \mathrm{E}[\log Z_t^2] < 0, \tag{13.10}$$

which is immediately implied if $a_1 < 1$ (since $\mathrm{E}[\log Z_t^2] \leq \log \mathrm{E}[Z_t^2] = 0$), but it is also satisfied under weaker conditions on a_1.

To obtain the moments of X_t^2 we use that it has the solution is

$$X_t^2 = Z_t^2 \left(a_0 + a_0 \sum_{k=1}^{\infty} a_1^k \prod_{j=1}^{k} Z_{t-j}^2 \right), \tag{13.11}$$

therefore taking expectations we have

$$\mathrm{E}[X_t^2] \;=\; \mathrm{E}[Z_t^2]\mathrm{E}\left(a_0 + a_0 \sum_{k=1}^{\infty} a_1^k \prod_{j=1}^{k} Z_{t-j}^2 \right) = a_0 + a_0 \sum_{k=1}^{\infty} a_1^k.$$

Thus $\mathrm{E}[X_t^2] < \infty$ if and only if $a_1 < 1$ (heuristically we can see this from $\mathrm{E}[X_t^2] = \mathrm{E}[Z_2^2](a_0 + a_1 \mathrm{E}[X_{t-1}^2])$).

By placing stricter conditions on a_1, namely $a_1 \mathrm{E}(Z_t^{2d})^{1/d} < 1$, we can show that $\mathrm{E}[X_t^{2d}] < \infty$ (note that this is an iff condition). To see why consider the simple case d is an integer, then by

using (13.11) we have

$$\begin{aligned} X_t^{2d} &\geq Z_t^{2d} a_0^d \sum_{k=1}^{\infty} a_1^{dk} \left(\prod_{j=1}^{k} Z_{t-j}^2 \right)^{2d} \\ \Rightarrow \mathrm{E}[X_t^{2d}] &\geq \mathrm{E}[Z_t^{2d}] a_0^d \sum_{k=1}^{\infty} a_1^{dk} \prod_{j=1}^{k} \mathrm{E}[Z_{t-j}^{2d}] = \mathrm{E}[Z_t^{2d}] a_0^d \sum_{k=1}^{\infty} a_1^{dk} \mathrm{E}[Z_t^{2d}]^k \\ &= \mathrm{E}[Z_t^{2d}] a_0^d \sum_{k=1}^{\infty} \left(a_1^d \mathrm{E}[Z_t^{2d}] \right)^k. \end{aligned}$$

It is immediately clear the above is only finite if $a_1 \mathrm{E}[Z_t^{2d}]^{1/d} < 1$.

The ARCH(p) model

We can generalize the above results to ARCH(p) processes (but to show existence of a solution we need to write the ARCH(p) process as a vector process similar to the Vector AR(1) representation of an AR(p) given in Section **??**). It can be shown that under sufficient conditions on the coefficients $\{a_j\}$ that the stationary, causal solution of the ARCH(p) model is

$$\begin{aligned} X_t^2 &= a_0 Z_t^2 + \sum_{k \geq 1} m_t(k) \qquad (13.12) \\ \text{where } m_t(k) &= \sum_{j_1, \ldots, j_k \geq 1} a_0 \Big(\prod_{r=1}^{k} a_{j_r} \Big) \prod_{r=0}^{k} Z_{t-\sum_{s=0}^{r} j_s}^2 \quad (j_0 = 0). \end{aligned}$$

The above solution belongs to a general class of functions called a Volterra expansion. We note that $\mathrm{E}[X_t^2] < \infty$ iff $\sum_{j=1}^{p} a_j < 1$.

13.3 The GARCH model

A possible drawback of the ARCH(p) model is that the conditional variance only depends on finite number of the past squared observations/log returns (in finance, the share price is often called the return). However, when fitting the model to the data, analogous to order selection of an autoregressive model (using, say, the AIC), often a large order p is selected. This suggests that the conditional variance should involve a large (infinite?) number of past terms. This observation motivated the GARCH model (first proposed in Bollerslev (1986) and Taylor (1986)), which in many respects is analogous to the ARMA. The conditional variance of the GARCH model is a

weighted average of the squared returns, the weights decline with the lag, but never go completely to zero. The GARCH class of models is a rather parsimonous class of models and is extremely popular in finance. The GARCH(p,q) model is defined as

$$X_t = \sigma_t Z_t \qquad \sigma_t^2 = a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2 + \sum_{i=1}^{q} b_i \sigma_{t-i}^2 \tag{13.13}$$

where Z_t are iid random variables where $\mathrm{E}(Z_t) = 0$ and $\mathrm{var}(Z_t) = 1$, $a_0 > 0$ and for $1 \leq j \leq p$ $a_j \geq 0$ and $1 \leq i \leq q$ $b_i \geq 0$.

Under the assumption that a causal solution with sufficient moments exist, the same properties defined for the ARCH(p) in Section 13.2.1 also apply to the GARCH(p,q) model.

It can be shown that under suitable conditions on $\{b_j\}$ that X_t satisfies an ARCH(∞) representation. Formally, we can write the conditional variance σ_t^2 (assuming that a stationarity solution exists) as

$$(1 - \sum_{i=1}^{q} b_i B^i)\sigma_t^2 = (a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2),$$

where B denotes the backshift notation defined in Chapter 4. Therefore if the roots of $b(z) = (1 - \sum_{i=1}^{q} b_i z^i)$ lie outside the unit circle (which is satisfied if $\sum_i b_i < 1$) then

$$\sigma_t^2 = \frac{1}{(1 - \sum_{j=1}^{q} b_j B^j)}(a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2) = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j X_{t-j}^2, \tag{13.14}$$

where a recursive equation for the derivation of α_j can be found in Berkes et al. (2003). In other words the GARCH(p,q) process can be written as a ARCH(∞) process. This is analogous to the invertibility representation given in Definition 4.5.2. This representation is useful when estimating the parameters of a GARCH process (see Berkes et al. (2003)) and also prediction. The expansion in (13.14) helps explain why the GARCH(p,q) process is so popular. As we stated at the start of this section, the conditional variance of the GARCH is a weighted average of the squared returns, the weights decline with the lag, but never go completely to zero, a property that is highly desirable.

Example 13.3.1 (Inverting the GARCH$(1,1)$) *If $b_1 < 1$, then we can write σ_t^2 as*

$$\sigma_t^2 = \left[\sum_{j=0}^{\infty} b^j B^j\right] \cdot \left[a_0 + a_1 X_{t-1}^2\right] = \frac{a_0}{1-b} + a_1 \sum_{j=0}^{\infty} b^j X_{t-1-j}^2.$$

This expansion offers us a clue as to why the GARCH$(1,1)$ is so popular in finance. In finance one important objective is to predict future volatility, this is the variance of say a stock tomorrow given past information. Using the GARCH model this is σ_t^2, which we see is

$$\sigma_t^2 \;=\; \frac{a_0}{1-b} + a_1 \sum_{j=0}^{\infty} b^j X_{t-1-j}^2.$$

This can be viewed as simply an exponentially weighted average of X_{t-j}^2. Some researchers argue that other models can lead to the same predictor of future volatility and there is nothing intrinsically specially about the GARCH process. We discuss this in more detail in Chapter 5.

In the following section we derive conditions for existence of the GARCH model and also it's moments.

13.3.1 Existence of a stationary solution of a GARCH$(1,1)$

We will focus on the GARCH$(1,1)$ model as this substantially simplifies the conditions. We recall the conditional variance of the GARCH$(1,1)$ can be written as

$$\sigma_t^2 = a_0 + a_1 X_{t-1}^2 + b_1 \sigma_{t-1}^2 = \left(a_1 Z_{t-1}^2 + b_1\right) \sigma_{t-1}^2 + a_0. \tag{13.15}$$

We observe that (13.15) can be written in the stochastic recurrence relation form given in (13.2) with $A_t = (a_1 Z_{t-1}^2 + b_1)$ and $B_t = a_0$. Therefore, by using Theorem 13.0.1, if $\mathrm{E}[\log(a_1 Z_{t-1}^2 + b_1)] < 0$, then σ_t^2 has the strictly stationary causal solution

$$\sigma_t^2 = a_0 + a_0 \sum_{k=1}^{\infty} \prod_{j=1}^{k} (a_1 Z_{t-j}^2 + b_1). \tag{13.16}$$

These conditions are relatively weak and depend on the distribution of Z_t. They are definitely satisfied if $a_1 + b_1 < 1$, since $\mathrm{E}[\log(a_1 Z_{t-1}^2 + b_1)] \leq \log \mathrm{E}[a_1 Z_{t-1}^2 + b_1] = \log(a_1 + b_1)$. However existence of a stationary solution does not require such a strong condition on the coefficients (and there can still exist a stationary solution if $a_1 + b_1 > 1$, so long as the distribution of Z_t^2 is such that $\mathrm{E}[\log(a_1 Z_t^2 + b_1)] < 0$).

By taking expectations of (13.16) we can see that

$$\mathrm{E}[X_t^2] = \mathrm{E}[\sigma_t^2] = a_0 + a_0 \sum_{k=1}^{\infty} \prod_{j=1}^{k} (a_1 + b_1) = a_0 + a_0 \sum_{k=1}^{\infty} (a_1 + b_1)^k.$$

Thus $\mathrm{E}[X_t^2] < \infty$ iff $a_1 + b_1 < 1$ (noting that a_1 and b_1 are both positive). Expanding on this argument, if $d > 1$ we can use Minkowski inequality to show

$$(\mathrm{E}[\sigma_t^{2d}])^{1/d} \leq a_0 + a_0 \sum_{k=1}^{\infty} (\mathrm{E}[\prod_{j=1}^{k} (a_1 Z_{t-j}^2 + b_1)]^d)^{1/d} \leq a_0 + a_0 \sum_{k=1}^{\infty} (\prod_{j=1}^{k} \mathrm{E}[(a_1 Z_{t-j}^2 + b_1)^d])^{1/d}.$$

Therefore, if $\mathrm{E}[(a_1 Z_{t-j}^2 + b_1)^d] < 1$, then $\mathrm{E}[X_t^{2d}] < \infty$. This is an iff condition, since from the definition in (13.15) we have

$$\mathrm{E}[\sigma_t^{2d}] = \mathrm{E}[\underbrace{a_0 + (a_1 Z_{t-1}^2 + b_1)\sigma_{t-1}^2}_{\text{every term is positive}}]^d \geq \mathrm{E}[(a_1 Z_{t-1}^2 + b_1)\sigma_{t-1}^2]^d = \mathrm{E}[(a_1 Z_{t-1}^2 + b_1)^d]\mathrm{E}[\sigma_{t-1}^{2d}],$$

since σ_{t-1}^2 has a causal solution, it is independent of Z_{t-1}^2. We observe that by stationarity and if $\mathrm{E}[\sigma_t^{2d}] < \infty$, then $\mathrm{E}[\sigma_t^{2d}] = \mathrm{E}[\sigma_{t-1}^{2d}]$. Thus the above inequality only holds if $\mathrm{E}[(a_1 Z_{t-1}^2 + b_1)^d] < 1$. Therefore, $\mathrm{E}[X_t^{2d}] < \infty$ iff $\mathrm{E}[(a_1 Z_{t-1}^2 + b_1)^d] < 1$.

Indeed in order for $\mathrm{E}[X_t^{2d}] < \infty$ a huge constraint needs to be placed on the parameter space of a_1 and b_1.

Exercise 13.1 *Suppose $\{Z_t\}$ are standard normal random variables. Find conditions on a_1 and b_1 such that $\mathrm{E}[X_t^4] < \infty$.*

The above results can be generalised to GARCH(p, q) model. Conditions for existence of a stationary solution hinge on the random matrix corresponding to the SRE representation of the GARCH model (see Bougerol and Picard (1992a) and Bougerol and Picard (1992b)), which are nearly impossible to verify. Sufficient and necessary conditions for both a stationary (causal) solution and second order stationarity ($\mathrm{E}[X_t^2] < \infty$) is $\sum_{j=1}^{p} a_j + \sum_{i=1}^{q} b_i < 1$. However, many econometricians believe this condition places an unreasonable constraint on the parameter space of $\{a_j\}$ and $\{b_j\}$. A large amount of research has been done on finding consistent parameter estimators under weaker conditions. Indeed, in the very interesting paper by Berkes et al. (2003) (see also Straumann (2005)) they derive consistent estimates of GARCH parameters on far milder set of conditions on $\{a_j\}$ and $\{b_i\}$ (which don't require $\mathrm{E}(X_t^2) < \infty$).

Definition 13.3.1 *The IGARCH model is a GARCH model where*

$$X_t = \sigma_t Z_t \qquad \sigma_t^2 = a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2 + \sum_{i=1}^{q} b_i \sigma_{t-i}^2 \tag{13.17}$$

where the coefficients are such that $\sum_{j=1}^{p} a_j + \sum_{i=1}^{q} b_i = 1$. *This is an example of a time series model which has a strictly stationary solution but it is not second order stationary.*

Exercise 13.2 *Simulate realisations of ARCH*(1) *and GARCH*$(1,1)$ *models. Simulate with both iid Gaussian and t-distribution errors* ($\{Z_t\}$ *where* $\mathrm{E}[Z_t^2] = 1$). *Remember to 'burn-in' each realisation.*

In all cases fix $a_0 > 0$. *Then*

(i) Simulate an ARCH(1) *with* $a_1 = 0.3$ *and* $a_1 = 0.9$.

(ii) Simulate a GARCH$(1,1)$ *with* $a_1 = 0.1$ *and* $b_1 = 0.85$, *and a GARCH*$(1,1)$ *with* $a_1 = 0.85$ *and* $b_1 = 0.1$. *Compare the two behaviours.*

13.3.2 Extensions of the GARCH model

One criticism of the GARCH model is that it is 'blind' to negative the sign of the return X_t. In other words, the conditional variance of X_t only takes into account the magnitude of X_t and does not depend on increases or a decreases in S_t (which corresponds to X_t being positive or negative). In contrast it is largely believed that the financial markets react differently to negative or positive X_t. The general view is that there is greater volatility/uncertainity/variation in the market when the price goes down.

This observation has motivated extensions to the GARCH, such as the EGARCH which take into account the sign of X_t. Deriving conditions for such a stationary solution to exist can be difficult task, and the reader is refered to Straumann (2005) and more the details.

Other extensions to the GARCH include an Autoregressive type model with GARCH innovations.

13.3.3 `R code`

`install.packages("tseries"), library("tseries")` recently there have been a new package developed `library("fGARCH")`.

13.4 Bilinear models

The Bilinear model was first proposed in Subba Rao (1977) and Granger and Andersen (1978) (see also Subba Rao (1981)). The general Bilinear (BL(p, q, r, s)) model is defined as

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{k=1}^{r} \sum_{k'=1}^{s} b_{k,k'} X_{t-k} \varepsilon_{t-k'},$$

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

To motivate the Bilinear model let us consider the simplest version of the model BL$(1, 0, 1, 1)$

$$X_t = \phi_1 X_{t-1} + b_{1,1} X_{t-1} \varepsilon_{t-1} + \varepsilon_t = [\phi_1 + b_{1,1} \varepsilon_{t-1}] X_{t-1} + \varepsilon_t. \tag{13.18}$$

Comparing (13.20) with the conditional variane of the GARCH$(1, 1)$ in (13.15) we see that they are very similar, the main differences are that (a) the bilinear model does not constrain the coefficients to be positive (whereas the conditional variance requires the coefficients to be positive) (b) the ε_{t-1} depends on X_{t-1}, whereas in the GARCH$(1, 1)$ Z_{t-1}^2 and σ_{t-1}^2 are independent coefficients and (c) the innovation in the GARCH$(1, 1)$ model is deterministic, whereas in the innovation in the Bilinear model is random. (b) and (c) makes the analysis of the Bilinear more complicated than the GARCH model. From model (13.20) we observe that when ε_{t-1} and X_{t-1} "couple" (both are large, mainly because ε_{t-1} is large) it leads to large burst in X_t. We observe this from the simulations below. Therefore this model has been used to model seismic activity etc.

13.4.1 Features of the Bilinear model

In this section we assume a causal, stationary solution of the bilinear model exists, the appropriate number of moments and that it is invertible in the sense that there exists a function g such that $\varepsilon_t = g(X_{t-1}, X_{t-2}, \ldots)$.

Under the assumption that the Bilinear process is invertible we can show that

$$\begin{aligned} \mathrm{E}[X_t | X_{t-1}, X_{t-2}, \ldots] &= \mathrm{E}[(\phi_1 + b_{1,1}\varepsilon_{t-1}) X_{t-1} | X_{t-1}, X_{t-2}, \ldots] + \mathrm{E}[\varepsilon_t | X_{t-1}, X_{t-2}, \ldots] \\ &= (\phi_1 + b_{1,1}\varepsilon_{t-1}) X_{t-1}, \end{aligned} \tag{13.19}$$

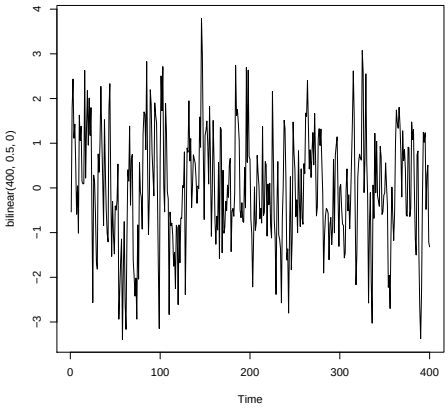
thus unlike the autoregressive model the conditional expectation of the X_t given the past is a nonlinear function of the past. It is this nonlinearity that gives rise to the spontaneous peaks that

we see a typical realisation.

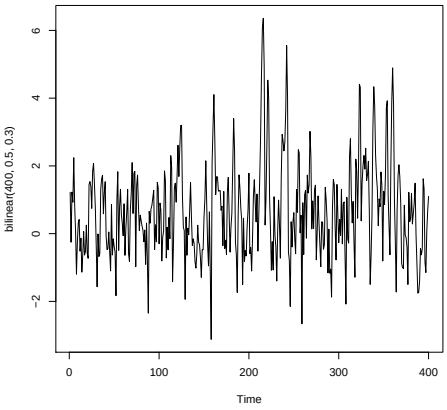
To see how the bilinear model was motivated in Figure 13.7 we give a plot of

$$X_t \quad = \quad \phi_1 X_{t-1} + b_{1,1} X_{t-1}\varepsilon_{t-1} + \varepsilon_t, \tag{13.20}$$

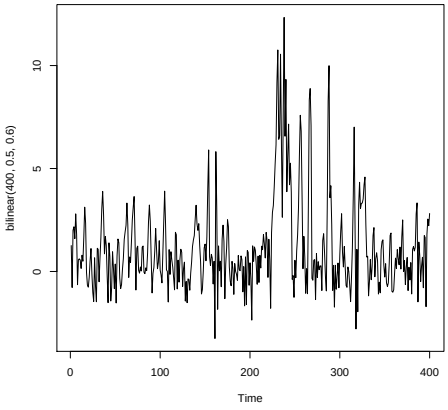
where $\phi_1 = 0.5$ and $b_{1,1} = 0, 0.35, 0.65$ and -0.65. and $\{\varepsilon_t\}$ are iid standard normal random variables. We observe that Figure 13.7a is a realisation from an AR(1) process and the subsequent plots are for different values of $b_{1,1}$. Figure 13.7a is quite 'regular', whereas the sudden bursts in activity become more pronounced as $b_{1,1}$ grows (see Figures 13.7b and 13.7c). In Figure 13.7d we give a plot realisation from a model where $b_{1,1}$ is negative and we see that the fluctation has changed direction.



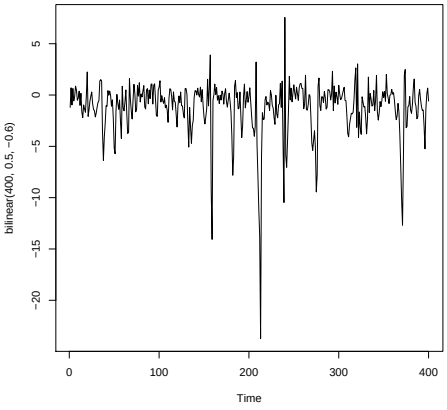
(a) $\phi_1 = 0.5$ and $b_{1,1} = 0$



(b) $\phi_1 = 0.5$ and $b_{1,1} = 0.35$



(c) $\phi_1 = 0.5$ and $b_{1,1} = 0.65$



(d) $\phi_1 = 0.5$ and $b_{1,1} = -0.65$

Figure 13.7: Realisations from different BL$(1, 0, 1, 1)$ models

Remark 13.4.1 (Markov Bilinear model) *Some authors define the* $BL(1,0,1,1)$ *as*

$$Y_t = \phi_1 Y_{t-1} + b_{1,1} Y_{t-1}\varepsilon_t + \varepsilon_t = [\phi_1 + b_{11}\varepsilon_t]Y_{t-1} + \varepsilon_t.$$

The fundamental difference between this model and (13.20) is that the multiplicative innovation (using ε_t *rather than* ε_{t-1}*) does not depend on* Y_{t-1}*. This means that* $\mathrm{E}[Y_t|Y_{t-1}, Y_{t-2}, \ldots] = \phi_1 Y_{t-1}$ *and the autocovariance function is the same as the autocovariance function of an* $AR(1)$ *model with the same AR parameter. Therefore, it is unclear the advantage of using this version of the model if the aim is to forecast, since the forecast of this model is the same as a forecast using the corresponding* $AR(1)$ *process* $X_t = \phi_1 X_{t-1} + \varepsilon_t$*. Forecasting with this model does not take into account its nonlinear behaviour.*

13.4.2 Solution of the Bilinear model

We observe that (13.20) can be written in the stochastic recurrence relation form given in (13.2) with $A_t = (\phi_1 + b_{11}\varepsilon_{t-1})$ and $B_t = a_0$. Therefore, by using Theorem 13.0.1, if $\mathrm{E}[\log(\phi_1 + b_{11}\varepsilon_{t-1})^2] < 0$ and $\mathrm{E}[\varepsilon_t] < \infty$, then X_t has the strictly stationary, causal solution

$$X_t = \sum_{k=1}^{\infty} \left[\prod_{j=1}^{k-1} (\phi_1 + b_{1,1}\varepsilon_{t-j}) \right] \cdot [(\phi_1 + b_{1,1}\varepsilon_{t-k})\varepsilon_{t-k}] + \varepsilon_t. \tag{13.21}$$

To show that it is second order stationary we require that $\mathrm{E}[X_t^2] < \infty$, which imposes additional conditions on the parameters. To derive conditions for $\mathrm{E}[X_t^2]$ we use (13.22) and the Minkowski inequality to give

$$\begin{aligned} (\mathrm{E}[X_t^2])^{1/2} &\leq \sum_{k=1}^{\infty} \mathrm{E}\left(\left[\prod_{j=1}^{k-1} (\phi_1 + b_{1,1}\varepsilon_{t-j}) \right]^2 \right)^{1/2} \cdot \left(\mathrm{E}\left[(\phi_1 + b_{11}\varepsilon_{t-k})\varepsilon_{t-k} \right]^2 \right)^{1/2} \\ &= \sum_{k=1}^{\infty} \prod_{j=1}^{k-1} \mathrm{E}\left([(\phi_1 + b_{1,1}\varepsilon_{t-j})]^2 \right)^{1/2} \cdot \left(\mathrm{E}\left[(\phi_1 + b_{1,1}\varepsilon_{t-k})\varepsilon_{t-k} \right]^2 \right)^{1/2}. \end{aligned} \tag{13.22}$$

Therefore if $\mathrm{E}[\varepsilon_t^4] < \infty$ and

$$\mathrm{E}\left[(\phi_1 + b_{1,1}\varepsilon_t) \right]^2 = \phi^2 + b_{11}^2 \mathrm{var}(\varepsilon_t) < 1,$$

then $\mathrm{E}[X_t^2] < \infty$ (note that the above equality is due to $\mathrm{E}[\varepsilon_t] = 0$).

Remark 13.4.2 (Inverting the Bilinear model) *We note that*

$$\varepsilon_t \;\; = \;\; -(bX_{t-1})\varepsilon_{t-1} + [X_t - \phi X_{t-1}],$$

thus by iterating backwards with respect to ε_{t-j} *we have*

$$\varepsilon_t = \sum_{j=0}^{\infty} \left((-b)^{j-1} \prod_{i=0}^{j} X_{t-1-j} \right) [X_{t-j} - \phi X_{t-j-1}].$$

This invertible representation is useful both in forecasting and estimation (see Section 7.10.3).

Exercise 13.3 *Simulate the* $BL(2,0,1,1)$ *model (using the* $AR(2)$ *parameters* $\phi_1 = 1.5$ *and* $\phi_2 = -0.75$*). Experiment with different parameters to give different types of behaviour.*

Exercise 13.4 *The random coefficient AR model is a nonlinear time series proposed by Barry Quinn (see Nicholls and Quinn (1982) and Aue et al. (2006)). The random coefficient* $AR(1)$ *model is defined as*

$$X_t = (\phi + \eta_t)X_{t-1} + \varepsilon_t$$

where $\{\varepsilon_t\}$ *and* $\{\eta_t\}$ *are iid random variables which are independent of each other.*

(i) State sufficient conditions which ensure that $\{X_t\}$ *has a strictly stationary solution.*

(ii) State conditions which ensure that $\{X_t\}$ *is second order stationary.*

(iii) Simulate from this model for different ϕ *and* $\text{var}[\eta_t]$.

13.4.3 R code

Code to simulate a $\text{BL}(1,0,1,1)$ model:

```
# Bilinear Simulation
# Bilinear(1,0,1,1) model, we use the first n0 observations are burn-in
# in order to get close to the stationary solution.
bilinear <- function(n,phi,b,n0=400) {
y <- rnorm(n+n0)
```

```
w <- rnorm(n + n0)
for (t in 2:(n+n0)) {
y[t] <- phi * y[t-1] + b * w[t-1] * y[t-1] + w[t]
}
return(y[(n0+1):(n0+n)])
}
```

13.5 Nonparametric time series models

Many researchers argue that fitting parametric models can lead to misspecification and argue that it may be more realistic to fit nonparametric or semi-parametric time series models instead. There exists several nonstationary and semi-parametric time series (see Fan and Yao (2003) and Douc et al. (2014) for a comprehensive summary), we give a few examples below. The most general nonparametric model is

$$X_t = m(X_{t-1}, \ldots, X_{t-p}, \varepsilon_t),$$

but this is so general it looses all meaning, especially if the need is to predict. A slight restriction is make the innovation term additive (see Jones (1978))

$$X_t = m(X_{t-1}, \ldots, X_{t-p}) + \varepsilon_t,$$

it is clear that for this model $\mathrm{E}[X_t|X_{t-1}, \ldots, X_{t-p}] = m(X_{t-1}, \ldots, X_{t-p})$. However this model has the distinct disadvantage that without placing any structure on $m(\cdot)$, for $p > 2$ nonparametric estimators of $m(\cdot)$ are lousy (as the suffer from the curse of dimensionality).

Thus such a generalisation renders the model useless. Instead semi-parametric approaches have been developed. Examples include the functional AR(p) model defined as

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

the semi-parametric AR(1) model

$$X_t = \phi X_{t-1} + \gamma(X_{t-1}) + \varepsilon_t,$$

the nonparametric ARCH(p)

$$X_t = \sigma_t Z_t \qquad \sigma_t^2 = a_0 + \sum_{j=1}^{p} a_j(X_{t-j}^2).$$

In the case of all these models it is not easy to establish conditions in which a stationary solution exists. More often then not, if conditions are established they are similar in spirit to those that are used in the parametric setting. For some details on the proof see Vogt (2013) (also here), who considers nonparametric and nonstationary models (note the nonstationarity he considers is when the covariance structure changes over time, not the unit root type). For example in the case of the the semi-parametric AR(1) model, a stationary causal solution exists if $|\phi + \gamma'(0)| < 1$.

Remark 13.5.1 *There are several other parametric nonlinear models that are of interest. A notable example, is the Markov Switching (Vector) Autorregressive model, which is popular in econometrics.*

This model generalises the classical autoregressive model in the sense that X_t has the presentation

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

where $S_t \in \{1, \ldots, M\}$ is a discrete Markov process that is unobserved and $\{\phi_j(s)\}_{j=1}^p$ are the $AR(p)$ parameters in state s. Note that for these models we are only interested in causal solutions (i.e. those which are generated from the past)

One fascinating aspect of this model, is that unlike the classical $AR(p)$ process, in order for X_t to have a well defined solution, the characteristic function corresponding to $\{\phi_j(s)\}_{j=1}^p$ need not have roots which lie outside the unit circle. If the switch into the state s is "short enough" the roots of $\{\phi_j(s)\}_{j=1}^p$ can lie within the unit circle (and a short explosion can occur), see **?** *for further details.*

Chapter 14

Consistency and and asymptotic normality of estimators

In the previous chapter we considered estimators of several different parameters. The hope is that as the sample size increases the estimator should get 'closer' to the parameter of interest. When we say closer we mean to converge. In the classical sense the sequence $\{x_k\}$ converges to x ($x_k \to x$), if $|x_k - x| \to 0$ as $k \to \infty$ (or for every $\varepsilon > 0$, there exists an n where for all $k > n$, $|x_k - x| < \varepsilon$). Of course the estimators we have considered are random, that is for every $\omega \in \Omega$ (set of all out comes) we have an different *estimate*. The natural question to ask is what does convergence mean for random sequences.

14.1 Modes of convergence

We start by defining different modes of convergence.

Definition 14.1.1 (Convergence)

- **Almost sure convergence** *We say that the sequence $\{X_t\}$ converges almost sure to μ, if there exists a set $M \subset \Omega$, such that $\mathbb{P}(M) = 1$ and for every $\omega \in N$ we have*

$$X_t(\omega) \to \mu.$$

In other words for every $\varepsilon > 0$, there exists an $N(\omega)$ such that

$$|X_t(\omega) - \mu| < \varepsilon, \tag{14.1}$$

for all $t > N(\omega)$. Note that the above definition is very close to classical convergence. We denote $X_t \to \mu$ almost surely, as $X_t \overset{a.s.}{\to} \mu$.

An equivalent definition, in terms of probabilities, is for every $\varepsilon > 0$ $X_t \overset{a.s.}{\to} \mu$ if

$$P(\omega; \cap_{m=1}^{\infty} \cup_{t=m}^{\infty} \{|X_t(\omega) - \mu| > \varepsilon\}) = 0.$$

It is worth considering briefly what $\cap_{m=1}^{\infty} \cup_{t=m}^{\infty} \{|X_t(\omega) - \mu| > \varepsilon\}$ means. If $\cap_{m=1}^{\infty} \cup_{t=m}^{\infty} \{|X_t(\omega) - \mu| > \varepsilon\} \neq \oslash$, then there exists an $\omega^* \in \cap_{m=1}^{\infty} \cup_{t=m}^{\infty} \{|X_t(\omega) - \mu| > \varepsilon\}$ such that for some infinite sequence $\{k_j\}$, we have $|X_{k_j}(\omega^*) - \mu| > \varepsilon$, this means $X_t(\omega^*)$ does not converge to μ. Now let $\cap_{m=1}^{\infty} \cup_{t=m}^{\infty} \{|X_t(\omega) - \mu| > \varepsilon\} = A$, if $P(A) = 0$, then for 'most' ω the sequence $\{X_t(\omega)\}$ converges.

- **Convergence in mean square**

  We say $X_t \to \mu$ in mean square (or L_2 convergence), if $\mathrm{E}(X_t - \mu)^2 \to 0$ as $t \to \infty$.

- **Convergence in probability**

  Convergence in probability cannot be stated in terms of realisations $X_t(\omega)$ but only in terms of probabilities. X_t is said to converge to μ in probability (written $X_t \overset{\mathcal{P}}{\to} \mu$) if

$$P(|X_t - \mu| > \varepsilon) \to 0, \quad t \to \infty.$$

  Often we write this as $|X_t - \mu| = o_p(1)$.

  If for any $\gamma \geq 1$ we have

$$\mathrm{E}(X_t - \mu)^{\gamma} \to 0 \quad t \to \infty,$$

  then it implies convergence in probability (to see this, use Markov's inequality).

- **Rates of convergence***:*

  (i) Suppose $a_t \to 0$ as $t \to \infty$. We say the stochastic process $\{X_t\}$ is $|X_t - \mu| = O_p(a_t)$,

if the sequence $\{a_t^{-1}|X_t - \mu|\}$ *is bounded in probability (this is defined below). We see from the definition of boundedness, that for all* t*, the distribution of* $a_t^{-1}|X_t - \mu|$ *should mainly lie within a certain interval.*

(ii) We say the stochastic process $\{X_t\}$ *is* $|X_t - \mu| = o_p(a_t)$*, if the sequence* $\{a_t^{-1}|X_t - \mu|\}$ *converges in probability to zero.*

Definition 14.1.2 (Boundedness) *(i)* **Almost surely bounded** *If the random variable* X *is almost surely bounded, then for a positive sequence* $\{e_k\}$*, such that* $e_k \to \infty$ *as* $k \to \infty$ *(typically* $e_k = 2^k$ *is used), we have*

$$P(\omega; \{\cup_{k=1}^{\infty}\{|X(\omega)| \leq e_k\}\}) = 1.$$

Usually to prove the above we consider the complement

$$P((\omega; \{\cup_{k=1}^{\infty}\{|X| \leq e_k\}\})^c) = 0.$$

Since $(\cup_{k=1}^{\infty}\{|X| \leq e_k\})^c = \cap_{k=1}^{\infty}\{|X| > e_k\} \subset \cap_{k=1}^{\infty} \cup_{m=k}^{\infty} \{|X| > e_k\}$*, to show the above we show*

$$P(\omega : \{\cap_{k=1}^{\infty} \cup_{m=k}^{\infty} \{|X(\omega)| > e_k\}\}) = 0. \tag{14.2}$$

We note that if $(\omega : \{\cap_{k=1}^{\infty} \cup_{m=k}^{\infty} \{|X(\omega)| > e_k\}\}) \neq \oslash$*, then there exists a* $\omega^* \in \Omega$ *and an infinite subsequence* k_j*, where* $|X(\omega^*)| > e_{k_j}$*, hence* $X(\omega^*)$ *is not bounded (since* $e_k \to \infty$*). To prove (14.2) we usually use the Borel Cantelli Lemma. This states that if* $\sum_{k=1}^{\infty} P(A_k) < \infty$*, the events* $\{A_k\}$ *occur only finitely often with probability one. Applying this to our case, if we can show that* $\sum_{m=1}^{\infty} P(\omega : \{|X(\omega)| > e_m|\}) < \infty$*, then* $\{|X(\omega)| > e_m|\}$ *happens only finitely often with probability one. Hence if* $\sum_{m=1}^{\infty} P(\omega : \{|X(\omega)| > e_m|\}) < \infty$*, then* $P(\omega : \{\cap_{k=1}^{\infty} \cup_{m=k}^{\infty} \{|X(\omega)| > e_k\}\}) = 0$ *and* X *is a bounded random variable.*

It is worth noting that often we choose the sequence $e_k = 2^k$*, in this case* $\sum_{m=1}^{\infty} P(\omega : \{|X(\omega)| > e_m|\}) = \sum_{m=1}^{\infty} P(\omega : \{\log|X(\omega)| > \log 2^k|\}) \leq C\mathrm{E}(\log|X|)$*. Hence if we can show that* $\mathrm{E}(\log|X|) < \infty$*, then* X *is bounded almost surely.*

b

(ii) **Sequences which are bounded in probability** *A sequence is bounded in probability,*

written $X_t = O_p(1)$, if for every $\varepsilon > 0$, there exists a $\delta(\varepsilon) < \infty$ such that $P(|X_t| \geq \delta(\varepsilon)) < \varepsilon$. Roughly speaking this means that the sequence is only extremely large with a very small probability. And as the 'largeness' grows the probability declines.

14.2 Sampling properties

Often we will estimate the parameters by maximising (or minimising) a criterion. Suppose we have the criterion $\mathcal{L}_n(a)$ (eg. likelihood, quasi-likelihood, Kullback-Leibler etc) we use as an estimator of a_0, $\hat{a}_n$ where

$$\hat{a}_n = \arg\max_{a \in \Theta} \mathcal{L}_n(a)$$

and Θ is the parameter space we do the maximisation (minimisation) over. Typically the true parameter a should maximise (minimise) the 'limiting' criterion $\mathcal{L}$.

If this is to be a good estimator, as the sample size grows the estimator should converge (in some sense) to the parameter we are interesting in estimating. As we discussed above, there are various modes in which we can measure this convergence (i) almost surely (ii) in probability and (iii) in mean squared error. Usually we show either (i) or (ii) (noting that (i) implies (ii)), in time series its usually quite difficult to show (iii).

Definition 14.2.1 *(i) An estimator $\hat{a}_n$ is said to be almost surely consistent estimator of a_0, if there exists a set $M \subset \Omega$, where $\mathbb{P}(M) = 1$ and for all $\omega \in M$ we have*

$$\hat{a}_n(\omega) \to a.$$

(ii) An estimator $\hat{a}_n$ is said to converge in probability to a_0, if for every $\delta > 0$

$$P(|\hat{a}_n - a| > \delta) \to 0 \quad T \to \infty.$$

To prove either (i) or (ii) usually involves verifying two main things, pointwise convergence and equicontinuity.

14.3 Showing almost sure convergence of an estimator

We now consider the general case where $\mathcal{L}_n(a)$ is a 'criterion' which we maximise. Let us suppose we can write $\mathcal{L}_n$ as

$$\mathcal{L}_n(a) = \frac{1}{n}\sum_{t=1}^{n} \ell_t(a), \tag{14.3}$$

where for each $a \in \Theta$, $\{\ell_t(a)\}_t$ is a ergodic sequence. Let

$$\mathcal{L}(a) = \mathrm{E}(\ell_t(a)), \tag{14.4}$$

we assume that $\mathcal{L}(a)$ is continuous and has a unique maximum in Θ. We define the estimator $\hat{\alpha}_n$ where $\hat{\alpha}_n = \arg\min_{a\in\Theta} \mathcal{L}_n(a)$.

Definition 14.3.1 (Uniform convergence) *$\mathcal{L}_n(a)$ is said to almost surely converge uniformly to $\mathcal{L}(a)$, if*

$$\sup_{a\in\Theta} |\mathcal{L}_n(a) - \mathcal{L}(a)| \stackrel{a.s.}{\to} 0.$$

In other words there exists a set $M \subset \Omega$ where $P(M) = 1$ and for every $\omega \in M$,

$$\sup_{a\in\Theta} |\mathcal{L}_n(\omega, a) - \mathcal{L}(a)| \to 0.$$

Theorem 14.3.1 (Consistency) *Suppose that $\hat{a}_n = \arg\max_{a\in\Theta} \mathcal{L}_n(a)$ and $a_0 = \arg\max_{a\in\Theta} \mathcal{L}(a)$ is the unique maximum. If $\sup_{a\in\Theta} |\mathcal{L}_n(a) - \mathcal{L}(a)| \stackrel{a.s.}{\to} 0$ as $n \to \infty$ and $\mathcal{L}(a)$ has a unique maximum. Then Then $\hat{a}_n \stackrel{a.s.}{\to} a_0$ as $n \to \infty$.*

PROOF. We note that by definition we have $\mathcal{L}_n(a_0) \leq \mathcal{L}_n(\hat{a}_n)$ and $\mathcal{L}(\hat{a}_n) \leq \mathcal{L}(a_0)$. Using this inequality we have

$$\mathcal{L}_n(a_0) - \mathcal{L}(a_0) \leq \mathcal{L}_n(\hat{a}_n) - \mathcal{L}(a_0) \leq \mathcal{L}_n(\hat{a}_n) - \mathcal{L}(\hat{a}_n).$$

Therefore from the above we have

$$|\mathcal{L}_n(\hat{a}_T) - \mathcal{L}(a_0)| \leq \max\{|\mathcal{L}_n(a_0) - \mathcal{L}(a_0)|, |\mathcal{L}_n(\hat{a}_T) - \mathcal{L}(\hat{a}_n)|\} \leq \sup_{a\in\Theta} |\mathcal{L}_n(a) - \mathcal{L}(a)|.$$

Hence since we have uniform converge we have $|\mathcal{L}_n(\hat{a}_n) - \mathcal{L}(a_0)| \stackrel{\text{a.s.}}{\to} 0$ as $n \to \infty$. Now since $\mathcal{L}(a)$ has a unique maximum, we see that $|\mathcal{L}_n(\hat{a}_n) - \mathcal{L}(a_0)| \stackrel{\text{a.s.}}{\to} 0$ implies $\hat{a}_n \stackrel{\text{a.s.}}{\to} a_0$. □

We note that directly establishing uniform convergence is not easy. Usually it is done by assuming the parameter space is compact and showing point wise convergence and stochastic equicontinuity, these three facts imply uniform convergence. Below we define stochastic equicontinuity and show consistency under these conditions.

Definition 14.3.2 *The sequence of stochastic functions $\{f_n(a)\}_n$ is said to be stochastically equicontinuous if there exists a set $M \in \Omega$ where $P(M) = 1$ and for every $\omega \in M$ and and $\varepsilon > 0$, there exists a δ and such that for every $\omega \in M$*

$$\sup_{|a_1 - a_2| \leq \delta} |f_n(\omega, a_1) - f_n(\omega, a_2)| \leq \varepsilon,$$

for all $n > N(\omega)$.

A sufficient condition for stochastic equicontinuity of $f_n(a)$ (which is usually used to prove equicontinuity), is that $f_n(a)$ is in some sense Lipschitz continuous. In other words,

$$\sup_{a_1, a_2 \in \Theta} |f_n(a_1) - f_n(a_2)| < K_n \|a_1 - a_2\|,$$

where k_n is a random variable which converges to a finite constant as $n \to \infty$ ($K_n \stackrel{a.s.}{\to} K_0$ as $n \to \infty$). To show that this implies equicontinuity we note that $K_n \stackrel{a.s.}{\to} K_0$ means that for every $\omega \in M$ ($P(M) = 1$) and $\gamma > 0$, we have $|K_n(\omega) - K_0| < \gamma$ for all $n > N(\omega)$. Therefore if we choose $\delta = \varepsilon/(K_0 + \gamma)$ we have

$$\sup_{|a_1 - a_2| \leq \varepsilon/(K_0 + \gamma)} |f_n(\omega, a_1) - f_n(\omega, a_2)| < \varepsilon,$$

for all $n > N(\omega)$.

In the following theorem we state sufficient conditions for almost sure uniform convergence. It is worth noting this is the Arzela-Ascoli theorem for random variables.

Theorem 14.3.2 (The stochastic Ascoli Lemma) *Suppose the parameter space Θ is compact, for every $a \in \Theta$ we have $\mathcal{L}_n(a) \stackrel{a.s.}{\to} \mathcal{L}(a)$ and $\mathcal{L}_n(a)$ is stochastic equicontinuous. Then $\sup_{a \in \Theta} |\mathcal{L}_n(a) - \mathcal{L}(a)| \stackrel{a.s.}{\to} 0$ as $n \to \infty$.*

We use the theorem below.

Corollary 14.3.1 *Suppose that* $\hat{a}_n = \arg\max_{a\in\Theta} \mathcal{L}_n(a)$ *and* $a_0 = \arg\max_{a\in\Theta} \mathcal{L}(a)$, *moreover* $\mathcal{L}(a)$ *has a unique maximum. If*

(i) We have point wise convergence, that is for every $a \in \Theta$ *we have* $\mathcal{L}_n(a) \stackrel{a.s.}{\to} \mathcal{L}(a)$.

(ii) The parameter space Θ *is compact.*

(iii) $\mathcal{L}_n(a)$ *is stochastic equicontinuous.*

then $\hat{a}_n \stackrel{a.s.}{\to} a_0$ *as* $n \to \infty$.

PROOF. By using Theorem 14.3.2 three assumptions imply that $|\sup_{\theta\in\Theta} ||\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| \to 0$, thus by using Theorem 14.3.1 we obtain the result.

We prove Theorem 14.3.2 in the section below, but it can be omitted on first reading.

14.3.1 Proof of Theorem 14.3.2 (The stochastic Ascoli theorem)

We now show that stochastic equicontinuity and almost pointwise convergence imply uniform convergence. We note that on its own, pointwise convergence is a much weaker condition than uniform convergence, since for pointwise convergence the rate of convergence can be different for each parameter.

Before we continue a few technical points. We recall that we are assuming almost pointwise convergence. This means for each parameter $a \in \Theta$ there exists a set $N_a \in \Omega$ (with $P(N_a) = 1$) such that for all $\omega \in N_a$ $\mathcal{L}_n(\omega, a) \to \mathcal{L}(a)$. In the following lemma we unify this set. That is show (using stochastic equicontinuity) that there exists a set $N \in \Omega$ (with $P(N) = 1$) such that for all $\omega \in N$ $\mathcal{L}_n(\omega, a) \to \mathcal{L}(a)$.

Lemma 14.3.1 *Suppose the sequence* $\{\mathcal{L}_n(a)\}_n$ *is stochastically equicontinuous and also pointwise convergent (that is* $\mathcal{L}_n(a)$ *converges almost surely to* $\mathcal{L}(a)$*), then there exists a set* $M \in \Omega$ *where* $P(\bar{M}) = 1$ *and for every* $\omega \in \bar{M}$ *and* $a \in \Theta$ *we have*

$$|\mathcal{L}_n(\omega, a) - \mathcal{L}(a)| \to 0.$$

PROOF. Enumerate all the rationals in the set Θ and call this sequence $\{a_i\}_i$. Since we have almost sure convergence, this implies for every a_i there exists a set M_{a_i} where $P(M_{a_i}) = 1$ and for every

$\omega \in M_{a_i}$ we have $|\mathcal{L}_T(\omega, a_i) - \mathcal{L}(a_i)| \to 0$. Define $M = \cap M_{a_i}$, since the number of sets is countable $P(M) = 1$ and for every $\omega \in M$ and a_i we have $\mathcal{L}_n(\omega, a_i) \to \mathcal{L}(a_i)$.

Since we have stochastic equicontinuity, there exists a set $\tilde{M}$ where $P(\tilde{M}) = 1$ and for every $\omega \in \tilde{M}$, $\{\mathcal{L}_n(\omega, \cdot)\}$ is equicontinuous. Let $\bar{M} = \tilde{M} \cap \{\cap M_{a_i}\}$, we will show that for all $a \in \Theta$ and $\omega \in \bar{M}$ we have $\mathcal{L}_n(\omega, a) \to \mathcal{L}(a)$. By stochastic equicontinuity for every $\omega \in \bar{M}$ and $\varepsilon/3 > 0$, there exists a $\delta > 0$ such that

$$\sup_{|b_1 - b_2| \leq \delta} |\mathcal{L}_n(\omega, b_1) - \mathcal{L}_n(\omega, b_2)| \leq \varepsilon/3, \tag{14.5}$$

for all $n > N(\omega)$. Furthermore by definition of $\bar{M}$ for every rational $a_j \in \Theta$ and $\omega \in \bar{N}$ we have

$$|\mathcal{L}_n(\omega, a_i) - \mathcal{L}(a_i)| \leq \varepsilon/3, \tag{14.6}$$

where $n > N'(\omega)$. Now for any given $a \in \Theta$, there exists a rational a_i such that $\|a - a_j\| \leq \delta$. Using this, (14.5) and (14.6) we have

$$|\mathcal{L}_n(\omega, a) - \mathcal{L}(a)| \leq |\mathcal{L}_n(\omega, a) - \mathcal{L}_n(\omega, a_i)| + |\mathcal{L}_n(\omega, a_i) - \mathcal{L}(a_i)| + |\mathcal{L}(a) - \mathcal{L}(a_i)| \leq \varepsilon,$$

for $n > \max(N(\omega), N'(\omega))$. To summarise for every $\omega \in \bar{M}$ and $a \in \Theta$, we have $|\mathcal{L}_n(\omega, a) - \mathcal{L}(a)| \to 0$. Hence we have pointwise covergence for every realisation in $\bar{M}$. □

We now show that equicontinuity implies uniform convergence.

Proof of Theorem 14.3.2. Using Lemma 14.3.1 we see that there exists a set $\bar{M} \in \Omega$ with $P(\bar{M}) = 1$, where $\mathcal{L}_n$ is equicontinuous and also pointwise convergent. We now show uniform convergence on this set. Choose $\varepsilon/3 > 0$ and let δ be such that for every $\omega \in \bar{M}$ we have

$$\sup_{|a_1 - a_2| \leq \delta} |\mathcal{L}_T(\omega, a_1) - \mathcal{L}_T(\omega, a_2)| \leq \varepsilon/3, \tag{14.7}$$

for all $n > n(\omega)$. Since Θ is compact it can be divided into a finite number of open sets. Construct the sets $\{O_i\}_{i=1}^p$, such that $\Theta \subset \cup_{i=1}^p O_i$ and $\sup_{x,y,i} \|x - y\| \leq \delta$. Let $\{a_i\}_{i=1}^p$ be such that $a_i \in O_i$. We note that for every $\omega \in \bar{M}$ we have $\mathcal{L}_n(\omega, a_i) \to \mathcal{L}(a_i)$, hence for every $\varepsilon/3$, there exists an $n_i(\omega)$ such that for all $n > n_i(\omega)$ we have $|\mathcal{L}_T(\omega, a_i) - \mathcal{L}(a_i)| \leq \varepsilon/3$. Therefore, since p is finite (due

to compactness), there exists a $\tilde{n}(\omega)$ such that

$$\max_{1\leq i\leq p} |\mathcal{L}_n(\omega, a_i) - \mathcal{L}(a_i)| \leq \varepsilon/3,$$

for all $n > \tilde{n}(\omega) = \max_{1\leq i\leq p}(n_i(\omega))$. For any $a \in \Theta$, choose the i, such that open set O_i such that $a \in O_i$. Using (14.7) we have

$$|\mathcal{L}_T(\omega, a) - \mathcal{L}_T(\omega, a_i)| \leq \varepsilon/3,$$

for all $n > n(\omega)$. Altogether this gives

$$|\mathcal{L}_T(\omega, a) - \mathcal{L}(a)| \leq |\mathcal{L}_T(\omega, a) - \mathcal{L}_T(\omega, a_i)| + |\mathcal{L}_T(\omega, a_i) - \mathcal{L}(a_i)| + |\mathcal{L}(a) - \mathcal{L}(a_i)| \leq \varepsilon,$$

for all $n \geq \max(n(\omega), \tilde{n}(\omega))$. We observe that $\max(n(\omega), \tilde{n}(\omega))$ and $\varepsilon/3$ does not depend on a, therefore for all $n \geq \max(n(\omega), \tilde{n}(\omega))$ and we have $\sup_a |\mathcal{L}_n(\omega, a) - \mathcal{L}(a)| < \varepsilon$. This gives for every $\omega \in \bar{M}$ ($\mathbb{P}(\bar{M}) = 1$), $\sup_a |\mathcal{L}_n(\omega, a) - \mathcal{L}(a)| \to 0$, thus we have almost sure uniform convergence. □

14.4 Toy Example: Almost sure convergence of the least squares estimator for an AR(p) process

In Chapter **??** we will consider the sampling properties of many of the estimators defined in Chapter 8. However to illustrate the consistency result above we apply it to the least squares estimator of the autoregressive parameters.

To simply notation we only consider estimator for $AR(1)$ models. Suppose that X_t satisfies $X_t = \phi X_{t-1} + \varepsilon_t$ (where $|\phi| < 1$). To estimate ϕ we use the least squares estimator defined below. Let

$$\mathcal{L}_n(a) = \frac{1}{n-1}\sum_{t=2}^{n}(X_t - aX_{t-1})^2, \tag{14.8}$$

we use $\hat{\phi}_n$ as an estimator of ϕ, where

$$\hat{\phi}_n = \arg\min_{a\in\Theta} \mathcal{L}_T(a), \tag{14.9}$$

where $\Theta = [-1, 1]$.

How can we show that this is consistent?

- In the case of least squares for AR processes, $\hat{a}_T$ has the explicit form

$$\hat{\phi}_n = \frac{\frac{1}{n-1}\sum_{t=2}^{n} X_t X_{t-1}}{\frac{1}{n-1}\sum_{t=1}^{T-1} X_t^2}.$$

  By just applying the ergodic theorem to the numerator and denominator we get $\hat{\phi}_n \stackrel{\text{a.s.}}{\to} \phi$.

  It is worth noting, that unlike the Yule-Walker estimator $\left|\frac{\frac{1}{n-1}\sum_{t=2}^{n} X_t X_{t-1}}{\frac{1}{n-1}\sum_{t=1}^{n-1} X_t^2}\right| < 1$ is not necessarily true.

- Here we will tackle the problem in a rather artifical way and assume that it does not have an explicit form and instead assume that $\hat{\phi}_n$ is obtained by minimising $\mathcal{L}_n(a)$ using a numerical routine.

- In order to derive the sampling properties of $\hat{\phi}_n$ we need to directly study the least squares criterion $\mathcal{L}_n(a)$. We will do this now in the least squares case.

We will first show almost sure convergence, which will involve repeated use of the ergodic theorem. We will then demonstrate how to show convergence in probability. We look at almost sure convergence as its easier to follow. Note that almost sure convergence implies convergence in probability (but the converse is not necessarily true).

The first thing to do it let

$$\ell_t(a) = (X_t - aX_{t-1})^2.$$

Since $\{X_t\}$ is an ergodic process (recall Example **??**(ii)) by using Theorem **??** we have for a, that $\{\ell_t(a)\}_t$ is an ergodic process. Therefore by using the ergodic theorem we have

$$\mathcal{L}_n(a) = \frac{1}{n-1}\sum_{t=2}^{n} \ell_t(a) \stackrel{\text{a.s.}}{\to} \mathrm{E}(\ell_0(a)).$$

In other words for every $a \in [-1, 1]$ we have that $\mathcal{L}_n(a) \stackrel{\text{a.s.}}{\to} \mathrm{E}(\ell_0(a))$ (almost sure pointwise convergence).

Since the parameter space $\Theta = [-1, 1]$ is compact and a is the unique minimum of $\ell(\cdot)$ in the

parameter space, all that remains is to show show stochastic equicontinuity. From this we deduce almost sure uniform convergence.

To show stochastic equicontinuity we expand $\mathcal{L}_T(a)$ and use the mean value theorem to obtain

$$\mathcal{L}_n(a_1) - \mathcal{L}_n(a_2) = \nabla \mathcal{L}_T(\bar{a})(a_1 - a_2), \tag{14.10}$$

where $\bar{a} \in [\min[a_1, a_2], \max[a_1, a_2]]$ and

$$\nabla \mathcal{L}_n(\bar{a}) = \frac{-2}{n-1} \sum_{t=2}^{n} X_{t-1}(X_t - \bar{a} X_{t-1}).$$

Because $\bar{a} \in [-1, 1]$ we have

$$|\nabla \mathcal{L}_n(\bar{a})| \leq \mathcal{D}_n, \text{ where } \mathcal{D}_n = \frac{2}{n-1} \sum_{t=2}^{n} (|X_{t-1} X_t| + X_{t-1}^2).$$

Since $\{X_t\}_t$ is an ergodic process, then $\{|X_{t-1} X_t| + X_{t-1}^2\}$ is an ergodic process. Therefore, if $\text{var}(\varepsilon_t) < \infty$, by using the ergodic theorem we have

$$\mathcal{D}_n \stackrel{\text{a.s.}}{\to} 2\text{E}(|X_{t-1} X_t| + X_{t-1}^2).$$

Let $\mathcal{D} := 2\text{E}(|X_{t-1} X_t| + X_{t-1}^2)$. Therefore there exists a set $M \in \Omega$, where $\mathbb{P}(M) = 1$ and for every $\omega \in M$ and $\varepsilon > 0$ we have

$$|\mathcal{D}_T(\omega) - \mathcal{D}| \leq \delta^*,$$

for all $n > N(\omega)$. Substituting the above into (14.10) we have

$$|\mathcal{L}_n(\omega, a_1) - \mathcal{L}_n(\omega, a_2)| \quad \leq \quad \mathcal{D}_n(\omega)|a_1 - a_2| \leq (\mathcal{D} + \delta^*)|a_1 - a_2|,$$

for all $n \geq N(\omega)$. Therefore for every $\varepsilon > 0$, there exists a $\delta := \varepsilon/(\mathcal{D} + \delta^*)$ such that

$$\sup_{|a_1 - a_2| \leq \varepsilon/(\mathcal{D} + \delta^*)} |\mathcal{L}_n(\omega, a_1) - \mathcal{L}_n(\omega, a_2)| \quad \leq \quad \varepsilon,$$

for all $n \geq N(\omega)$. Since this is true for all $\omega \in M$ we see that $\{\mathcal{L}_n(a)\}$ is stochastically equicontinuous.

Theorem 14.4.1 *Let $\hat{\phi}_n$ be defined as in (14.9). Then we have $\hat{\phi}_n \stackrel{a.s.}{\to} \phi$.*

PROOF. Since $\{\mathcal{L}_n(a)\}$ is almost sure equicontinuous, the parameter space $[-1,1]$ is compact and we have pointwise convergence of $\mathcal{L}_n(a) \stackrel{\text{a.s.}}{\to} \mathcal{L}(a)$, by using Theorem 14.3.1 we have that $\hat{\phi}_n \stackrel{\text{a.s.}}{\to} a$, where $a = \min_{a\in\Theta} \mathcal{L}(a)$. Finally we need to show that $a = \phi$. Since

$$\mathcal{L}(a) = \mathrm{E}(\ell_0(a)) = -\mathrm{E}(X_1 - aX_0)^2,$$

we see by differentiating $\mathcal{L}(a)$ with respect to a, that it is minimised at $a = \mathrm{E}(X_0X_1)/\mathrm{E}(X_0^2)$, hence $a = \mathrm{E}(X_0X_1)/\mathrm{E}(X_0^2)$. To show that this is ϕ, we note that by the Yule-Walker equations

$$X_t = \phi X_{t-1} + \epsilon_t \quad \Rightarrow \quad \mathrm{E}(X_tX_{t-1}) = \phi\mathrm{E}(X_{t-1}^2) + \underbrace{\mathrm{E}(\epsilon_tX_{t-1})}_{=0}.$$

Therefore $\phi = \mathrm{E}(X_0X_1)/\mathrm{E}(X_0^2)$, hence $\hat{\phi}_n \stackrel{\text{a.s.}}{\to} \phi$. □

We note that by using a very similar methods we can show strong consistency of the least squares estimator of the parameters in an AR(p) model.

14.5 Convergence in probability of an estimator

We described above almost sure (strong) consistency ($\hat{a}_T \stackrel{\text{a.s.}}{\to} a_0$). Sometimes its not possible to show strong consistency (eg. when ergodicity cannot be verified). As an alternative, weak consistency where $\hat{a}_T \stackrel{\mathcal{P}}{\to} a_0$ (convergence in probability), is shown. This requires a weaker set of conditions, which we now describe:

(i) The parameter space Θ should be compact.

(ii) Probability pointwise convergence: for every $a \in \Theta$ $\mathcal{L}_n(a) \stackrel{\mathcal{P}}{\to} \mathcal{L}(a)$.

(iii) The sequence $\{\mathcal{L}_n(a)\}$ is equicontinuous in probability. That is for every $\epsilon > 0$ and $\eta > 0$ there exists a δ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{|a_1-a_2|\leq\delta} |\mathcal{L}_n(a_1) - \mathcal{L}_n(a_2)| > \epsilon\right) < \eta. \tag{14.11}$$

If the above conditions are satisified we have $\hat{a}_T \stackrel{\mathcal{P}}{\to} a_0$.

Verifying conditions (ii) and (iii) may look a little daunting but by using Chebyshev's (or Markov's) inequality it can be quite straightforward. For example if we can show that for every $a \in \Theta$

$$\mathrm{E}(\mathcal{L}_n(a) - \mathcal{L}(a))^2 \to 0 \quad T \to \infty.$$

Therefore by applying Chebyshev's inequality we have for every $\varepsilon > 0$ that

$$P(|\mathcal{L}_n(a) - \mathcal{L}(a)| > \varepsilon) \leq \frac{\mathrm{E}(\mathcal{L}_n(a) - \mathcal{L}(a))^2}{\varepsilon^2} \to 0 \quad T \to \infty.$$

Thus for every $a \in \Theta$ we have $\mathcal{L}_n(a) \xrightarrow{\mathcal{P}} \mathcal{L}(a)$.

To show (iii) we often use the mean value theorem $\mathcal{L}_n(a)$. Using the mean value theorem we have

$$|\mathcal{L}_n(a_1) - \mathcal{L}_n(a_2)| \leq \sup_a \|\nabla_a \mathcal{L}_n(a)\|_2 \|a_1 - a_2\|.$$

Now if we can show that $\sup_n \mathrm{E} \sup_a \|\nabla_a \mathcal{L}_n(a)\|_2 < \infty$ (in other words it is uniformly bounded in probability over n) then we have the result. To see this observe that

$$\begin{aligned} \mathbb{P}\left(\sup_{|a_1 - a_2| \leq \delta} |\mathcal{L}_n(a_1) - \mathcal{L}_n(a_2)| > \epsilon \right) &\leq \mathbb{P}\left(\sup_{a \in \Omega} \|\nabla_a \mathcal{L}_n(a)\|_2 |a_1 - a_2| > \epsilon \right) \\ &\leq \frac{\sup_n \mathrm{E}(|a_1 - a_2| \sup_{a \in \Omega} \|\nabla_a \mathcal{L}_n(a)\|_2)}{\epsilon}. \end{aligned}$$

Therefore by a careful choice of $\delta > 0$ we see that (14.11) is satisfied (and we have equicontinuity in probability).

14.6 Asymptotic normality of an estimator

Once consistency of an estimator has been shown this paves the way to showing normality. To make the derivations simple we will assume that θ is univariate (this allows to easily use Taylor expansion) . We will assume that that the third derivative of the contrast function, $\mathcal{L}_n(\theta)$, exists, its expectation is bounded and it's variance converges to zero as $n \to \infty$. If this is the case we have have the following result

Lemma 14.6.1 *Suppose that the third derivative of the contrast function $\mathcal{L}_n(\theta)$ exists, for $k = 0, 1, 2$ $\mathrm{E}(\frac{\partial^k \mathcal{L}_n(\theta)}{\partial theta^k}) = \frac{\partial^k \mathcal{L}}{\partial \theta^k}$ and $\mathrm{var}(\frac{\partial^k \mathcal{L}_n(\theta)}{\partial theta^k}) \to 0$ as $n \to \infty$ and $\frac{\partial^3 \mathcal{L}_n(\theta)}{\partial theta^3}$ is bounded by a random variable Z_n which is independent of n where $\mathrm{E}(Z_n) < \infty$ and $\mathrm{var}(Z_n) \to 0$. Then we have*

$$(\hat{\theta}_n - \theta_0) = V(\theta)^{-1} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}_{\theta=\theta_0} + o_p(1) \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}_{\theta=\theta_0},$$

where $V(\theta_0) = \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}_{\theta_0}$.

PROOF. By the mean value theorem we have

$$\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}_{\theta=\theta_0} = \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}_{\theta=\hat{\theta}_n} - (\hat{\theta}_n - \theta_0) \frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}_{\theta=\bar{\theta}_n} = -(\hat{\theta}_n - \theta_0) \frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}_{\theta=\bar{\theta}_n} \quad (14.12)$$

where $\bar{\theta}_n$ lies between θ_0 and $\hat{\theta}_n$. We first study $\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}_{\theta=\bar{\theta}_n}$. By using the man value theorem we have

$$\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}_{\theta=\bar{\theta}_n} = \frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}_{\theta_0} + (\bar{\theta}_n - \theta_0) \frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}_{\theta=\tilde{\theta}_n}$$

where $\tilde{\theta}_n$ lies between θ_0 and $\bar{\theta}_n$. Since $\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}_{\theta_0} \to \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}_{\theta_0} = V(\theta_0)$, under the stated assumptions we have

$$\Big| \frac{\partial^2 \mathcal{L}}{\partial \theta^2}_{\theta=\bar{\theta}_n} - V(\theta_0) \Big| \leq |\bar{\theta}_n - \theta_0| \Big| \frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}_{\theta=\tilde{\theta}_n} \Big| \leq |\bar{\theta}_n - \theta_0| W_n.$$

Therefore, by consistency of the estimator it is clear that $\frac{\partial^2 \mathcal{L}}{\partial \theta^2}_{\theta=\bar{\theta}_n} \xrightarrow{\mathcal{P}} V(\theta_0)$. Substituting this into (14.12) we have

$$\frac{\partial \mathcal{L}}{\partial \theta}_{\theta=\theta_0} = -(\hat{\theta}_n - \theta_0)(V(\theta_0) + o(1)),$$

since $V(\theta_0)$ is bounded away from zero we have $[\frac{\partial^2 \mathcal{L}}{\partial \theta^2}_{\theta=\bar{\theta}_n}]^{-1} = V(\theta_0)^{-1} + o_p(1)$ and we obtain the desired result. □

The above result means that the distribution of $(\hat{\theta}_n - \theta_0)$ is determined by $\frac{\partial \mathcal{L}}{\partial \theta}_{\theta=\theta_0}$. In the following section we show to show asymptotic normality of $\frac{\partial \mathcal{L}}{\partial \theta}_{\theta=\theta_0}$.

14.6.1 Martingale central limit theorem

Remark 14.6.1 *We recall that*

$$(\hat{\phi}_n - \phi) = -\left(\nabla^2 \mathcal{L}_n\right)^{-1} \nabla \mathcal{L}_n(\phi) = \frac{\frac{-2}{n-1}\sum_{t=2}^n \varepsilon_t X_{t-1}}{\frac{2}{n-1}\sum_{t=2}^n X_{t-1}^2}, \tag{14.13}$$

and that $\mathrm{var}(\frac{-2}{n-1}\sum_{t=2}^n \varepsilon_t X_{t-1}) = \frac{-2}{n-1}\sum_{t=2}^n \mathrm{var}(\varepsilon_t X_{t-1}) = O(\frac{1}{n})$. *This implies*

$$(\hat{\phi}_n - \phi) = O_p(n^{-1/2}).$$

Indeed the results also holds almost surely

$$(\hat{\phi}_n - \phi) = O(n^{-1/2}). \tag{14.14}$$

The same result is true for autoregressive processes of arbitrary finite order. That is

$$\sqrt{n}(\hat{\underline{\phi}}_n - \underline{\phi}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{E}(\Gamma_p)^{-1}\sigma^2). \tag{14.15}$$

14.6.2 Example: Asymptotic normality of the weighted periodogram

Previously we have discussed the weight peiodogram, here we show normality of it, in the case that the time series X_t is zero mean linear time series (has the representation $X_t = \sum_j \psi_j \varepsilon_{t-j}$). Recalling Lemma 11.2.2 we have

$$\begin{aligned} A(\phi, I_n) &= \frac{1}{n}\sum_{k=1}^n \phi(\omega_k) I_n(\omega_k) \\ &= \frac{1}{n}\sum_{k=1}^n \phi(\omega_k)|A(\omega_k)^2| I_\varepsilon(\omega_k) + o(\frac{1}{n}). \end{aligned}$$

Therefore we will show asymptotic normality of $\frac{1}{n}\sum_{k=1}^n \phi(\omega_k)|A(\omega_k)^2| I_\varepsilon(\omega_k)$, which will give asymptotic normality of $A(\phi, I_n)$. Expanding $|I_\varepsilon(\omega_k)$ and substituting this into $\frac{1}{n}\sum_{k=1}^n \phi(\omega_k)|A(\omega_k)^2| I_\varepsilon(\omega_k)$ gives

$$\frac{1}{n}\sum_{k=1}^n \phi(\omega_k)|A(\omega_k)^2| I_\varepsilon(\omega_k) = \frac{1}{n}\sum_{t,\tau=1}^n \varepsilon_t\varepsilon_\tau \frac{1}{n}\sum_{k=1}^n \phi(\omega_k)|A(\omega_k)^2 \exp(i\omega_k(t-\tau)) = \frac{1}{n}\sum_{t,\tau=1}^n \varepsilon_t\varepsilon_\tau g_n(t-\tau)$$

where

$$g_n(t-\tau) = \frac{1}{n}\sum_{k=1}^{n} \phi(\omega_k)|A(\omega_k)^2 \exp(i\omega_k(t-\tau)) = \frac{1}{2\pi}\int_0^{2\pi} \phi(\omega)|A(\omega)|^2 \exp(i\omega(t-\tau))d\omega + O(\frac{1}{n^2}),$$

(the rate for the derivative exchange is based on assuming that the second derivatives of $A(\omega)$ and ϕ exist and $\phi(0) = \phi(2\pi)$). We can rewrite $\frac{1}{n}\sum_{t,\tau=1}^{n} \varepsilon_t\varepsilon_\tau g_n(t-\tau)$ as

$$\begin{aligned} & \frac{1}{n}\sum_{t,\tau=1}^{n} [\varepsilon_t\varepsilon_\tau - \mathrm{E}(\varepsilon_t\varepsilon_\tau)]g_n(t-\tau) \\ = & \frac{1}{n}\sum_{t=1}^{n} \Big([(\varepsilon_t^2 - \mathrm{E}(\varepsilon_t^2)]g_n(0) + \varepsilon_t\big(\sum_{\tau<t} \varepsilon_\tau[g_n(t-\tau) - g_n(\tau-t)]\big)\Big) \\ := & \frac{1}{n}\sum_{t=1}^{n} Z_{t,n} \end{aligned}$$

where it is straightforward to show that $\{Z_{t,n}\}$ are the sum of martingale differences. Thus we can show that

$$\frac{1}{\sqrt{n}}\sum_{t,\tau=1}^{n} \varepsilon_t\varepsilon_\tau g_n(t-\tau) - \mathrm{E}\big(\frac{1}{\sqrt{n}}\sum_{t,\tau=1}^{n} \varepsilon_t\varepsilon_\tau g_n(t-\tau)\big) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_{t,n}$$

satisfies the conditions of the martingale central limit theorem, which gives asymptotic normality of $\frac{1}{n}\sum_{t,\tau=1}^{n} \varepsilon_t\varepsilon_\tau g_n(t-\tau)$ and thus $A(\phi, I_n)$.

In the remainder of this chapter we obtain the sampling properties of the ARMA estimators defined in Sections 9.2.1 and 9.2.5.

14.7 Asymptotic properties of the Hannan and Rissanen estimation method

In this section we will derive the sampling properties of the Hannan-Rissanen estimator. We will obtain an almost sure rate of convergence (this will be the only estimator where we obtain an almost sure rate). Typically obtaining only sure rates can be more difficult than obtaining probabilistic rates, moreover the rates can be different (worse in the almost sure case). We now illustrate why that is with a small example. Suppose $\{X_t\}$ are iid random variables with mean zero and variance

one. Let $S_n = \sum_{t=1}^n X_t$. It can easily be shown that

$$\text{var}(S_n) = \frac{1}{n} \text{ therefore } S_n = O_p(\frac{1}{\sqrt{n}}). \tag{14.16}$$

However, from the law of iterated logarithm we have for any $\varepsilon > 0$

$$P(S_n \geq (1+\varepsilon)\sqrt{2n\log\log n} \text{ infinitely often}) = 0 P(S_n \geq (1-\varepsilon)\sqrt{2n\log\log n} \text{ infinitely often}) = 1 \tag{14.17}$$

Comparing (14.16) and (14.17) we see that for any given trajectory (realisation) most of the time $\frac{1}{n}S_n$ will be within the $O(\frac{1}{\sqrt{n}})$ bound but there will be excursions above when it to the $O(\frac{\log\log n}{\sqrt{n}}$ bound. In other words we cannot say that $\frac{1}{n}S_n = (\frac{1}{\sqrt{n}})$ almost surely, but we can say that This basically means that

$$\frac{1}{n}S_n = O(\frac{\sqrt{2\log\log n}}{\sqrt{n}}) \text{ almost surely.}$$

Hence the probabilistic and the almost sure rates are (slightly) different. Given this result is true for the average of iid random variables, it is likely that similar results will hold true for various estimators.

In this section we derive an almost sure rate for Hannan-Rissanen estimator, this rate will be determined by a few factors (a) an almost sure bound similar to the one derived above (b) the increasing number of parameters p_n (c) the bias due to estimating only a finite number of parameters when there are an infinite number in the model.

We first recall the algorithm:

(i) Use least squares to estimate $\{b_j\}_{j=1}^{p_n}$ and define

$$\hat{\mathbf{b}}_n = \hat{R}_n^{-1}\hat{\mathbf{r}}_n, \tag{14.18}$$

where $\hat{\mathbf{b}}_n' = (\hat{b}_{1,n}, \ldots, \hat{b}_{p_n,n})$,

$$\hat{R}_n = \sum_{t=p_n+1}^{n} \mathbf{X}_{t-1}\mathbf{X}_{t-1}' \quad \hat{\mathbf{r}}_n = \sum_{t=p_n+1}^{T} X_t\mathbf{X}_{t-1}$$

and $\mathbf{X}_{t-1}' = (X_{t-1}, \ldots, X_{t-p_n})$.

(ii) Estimate the residuals with

$$\tilde{\varepsilon}_t = X_t - \sum_{j=1}^{p_n} \hat{b}_{j,n} X_{t-j}.$$

(iii) Now use as estimates of $\underline{\phi}_0$ and $\underline{\theta}_0$ $\tilde{\underline{\phi}}_n, \tilde{\underline{\theta}}_n$ where

$$\tilde{\underline{\phi}}_n, \tilde{\underline{\theta}}_n = \arg\min \sum_{t=p_n+1}^{n} (X_t - \sum_{j=1}^{p} \phi_j X_{t-j} - \sum_{i=1}^{q} \theta_i \tilde{\varepsilon}_{t-i})^2. \tag{14.19}$$

We note that the above can easily be minimised. In fact

$$(\tilde{\underline{\phi}}_n, \tilde{\underline{\theta}}_n) = \tilde{\mathcal{R}}_n^{-1} \tilde{\mathbf{s}}_n$$

where

$$\tilde{\mathcal{R}}_n = \frac{1}{n} \sum_{t=p_n+1}^{n} \tilde{\mathbf{Y}}_t \tilde{\mathbf{Y}}_t \quad \tilde{\mathbf{s}}_n = \frac{1}{T} \sum_{t=p_n+1}^{n} \tilde{\mathbf{Y}}_t X_t,$$

$\tilde{\mathbf{Y}}_t' = (X_{t-1}, \ldots, X_{t-p}, \tilde{\varepsilon}_{t-1}, \ldots, \tilde{\varepsilon}_{t-q})$. Let $\tilde{\varphi}_n = (\tilde{\phi}_n, \tilde{\theta}_n)$.

We observe that in the second stage of the scheme where the estimation of the ARMA parameters are done, it is important to show that the empirical residuals are close to the true residuals. That is $\tilde{\varepsilon}_t = \varepsilon_t + o(1)$. We observe that from the definition of $\tilde{\varepsilon}_t$, this depends on the rate of convergence of the AR estimators $\hat{b}_{j,n}$

$$\begin{aligned} \tilde{\varepsilon}_t &= X_t - \sum_{j=1}^{p_n} \hat{b}_{j,n} X_{t-j} \\ &= \varepsilon_t + \sum_{j=1}^{p_n} (\hat{b}_{j,n} - b_j) X_{t-j} - \sum_{j=p_n+1}^{\infty} b_j X_{t-j}. \end{aligned} \tag{14.20}$$

Hence

$$\big|\hat{\varepsilon}_t - \varepsilon_t\big| \leq \big| \sum_{j=1}^{p_n} (\hat{b}_{j,n} - b_j) X_{t-j} \big| + \big| \sum_{j=p_n+1}^{\infty} b_j X_{t-j} \big|. \tag{14.21}$$

Therefore to study the asymptotic properties of $\tilde{\boldsymbol{\varphi}} = \tilde{\underline{\phi}}_n, \tilde{\underline{\theta}}_n$ we need to

- Obtain a rate of convergence for $\sup_j |\hat{b}_{j,n} - b_j|$.

- Obtain a rate for $|\hat{\varepsilon}_t - \varepsilon_t|$.
- Use the above to obtain a rate for $\tilde{\boldsymbol{\varphi}}_n = (\hat{\underline{\phi}}_n, \hat{\underline{\theta}}_n)$.

We first want to obtain the uniform rate of convergence for $\sup_j |\hat{b}_{j,n} - b_j|$. Deriving this is technically quite challanging. We state the rate in the following theorem, an outline of the proof can be found in Section 14.7.1. The proofs uses results from mixingale theory which can be found in Chapter B.

Theorem 14.7.1 *Suppose that $\{X_t\}$ is from an ARMA process where the roots of the true characteristic polynomials $\phi(z)$ and $\theta(z)$ both have absolute value greater than $1+\delta$. Let $\hat{\mathbf{b}}_n$ be defined as in (14.18), then we have almost surely*

$$\|\hat{\mathbf{b}}_n - \mathbf{b}_n\|_2 = O\left(p_n^2\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}} + \frac{p_n^3}{n} + p_n\rho^{p_n}\right)$$

for any $\gamma > 0$.

PROOF. See Section 14.7.1.

Corollary 14.7.1 *Suppose the conditions in Theorem 14.7.1 are satisfied. Then we have*

$$\left|\tilde{\varepsilon}_t - \varepsilon_t\right| \leq p_n \max_{1\leq j\leq p_n} |\hat{b}_{j,n} - b_j| Z_{t,p_n} + K\rho^{p_n} Y_{t-p_n}, \tag{14.22}$$

where $Z_{t,p_n} = \frac{1}{p_n}\sum_{t=1}^{p_n} |X_{t-j}|$ and $Y_t = \sum_{t=1}^{p_n} \rho^j |X_t|$,

$$\frac{1}{n}\sum_{t=p_n+1}^{n} \left|\tilde{\varepsilon}_{t-i}X_{t-j} - \varepsilon_{t-i}X_{t-j}\right| = O(p_n Q(n) + \rho^{p_n}) \tag{14.23}$$

$$\frac{1}{n}\sum_{t=p_n+1}^{n} \left|\tilde{\varepsilon}_{t-i}\tilde{\varepsilon}_{t-j} - \varepsilon_{t-i}\tilde{\varepsilon}_{t-j}\right| = O(p_n Q(n) + \rho^{p_n}) \tag{14.24}$$

where $Q(n) = p_n^2\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}} + \frac{p_n^3}{n} + p_n\rho^{p_n}$.

PROOF. Using (14.21) we immediately obtain (14.22).

To obtain (14.23) we use (14.21) to obtain

$$\begin{aligned}
& \frac{1}{n}\sum_{t=p_n+1}^{n}\left|\tilde{\varepsilon}_{t-i}X_{t-j}-\varepsilon_{t-i}X_{t-j}\right| \leq \frac{1}{n}\sum_{t=p_n+1}^{n}|X_{t-j}|\left|\tilde{\varepsilon}_{t-i}-\varepsilon_{t-i}\right| \\
\leq & O(p_nQ(n))\frac{1}{n}\sum_{t=p_n+1}^{n}|X_t||Z_{t,p_n}|+O(\rho^{p_n})\frac{1}{n}\sum_{t=p_n+1}^{n}|X_t||Y_{t-p_n}| \\
= & O(p_nQ(n)+\rho^{p_n}).
\end{aligned}$$

To prove (14.24) we use a similar method, hence we omit the details. □

We apply the above result in the theorem below.

Theorem 14.7.2 *Suppose the assumptions in Theorem 14.7.1 are satisfied. Then*

$$\left\|\tilde{\boldsymbol{\varphi}}_n-\boldsymbol{\varphi}_0\right\|_2 = O\left(p_n^3\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}}+\frac{p_n^4}{n}+p_n^2\rho^{p_n}\right).$$

for any $\gamma>0$, where $\tilde{\boldsymbol{\varphi}}_n=(\tilde{\underline{\phi}}_n,\tilde{\underline{\theta}}_n)$ and $\boldsymbol{\varphi}_0=(\underline{\phi}_0,\underline{\theta}_0)$.

PROOF. We note from the definition of $\tilde{\boldsymbol{\varphi}}_n$ that

$$\left(\tilde{\boldsymbol{\varphi}}_n-\boldsymbol{\varphi}_0\right) = \tilde{\mathcal{R}}_n^{-1}\left(\tilde{\mathbf{s}}_n-\tilde{\mathcal{R}}_n\tilde{\boldsymbol{\varphi}}_0\right).$$

Now in the $\tilde{\mathcal{R}}_n$ and $\tilde{\mathbf{s}}_n$ we replace the estimated residuals $\tilde{\varepsilon}_n$ with the true unobserved residuals. This gives us

$$\left(\tilde{\boldsymbol{\varphi}}_n-\boldsymbol{\varphi}_0\right) = \mathcal{R}_n^{-1}\left(\mathbf{s}_n-\mathcal{R}_n\boldsymbol{\varphi}_0\right)+\left(\mathcal{R}_n^{-1}\mathbf{s}_n-\tilde{\mathcal{R}}_n^{-1}\tilde{\mathbf{s}}_n\right) \tag{14.25}$$

$$\mathcal{R}_n=\frac{1}{n}\sum_{t=\max(p,q)}^{n}\mathbf{Y}_t\mathbf{Y}_t \quad \mathbf{s}_n=\frac{1}{n}\sum_{t=\max(p,q)}^{n}\mathbf{Y}_tX_t,$$

$\mathbf{Y}_t'=(X_{t-1},\ldots,X_{t-p},\varepsilon_{t-1},\ldots,\varepsilon_{t-q})$ (recalling that $\mathbf{Y}_t'=(X_{t-1},\ldots,X_{t-p},\tilde{\varepsilon}_{t-1},\ldots,\tilde{\varepsilon}_{t-q})$. The error term is

$$(\mathcal{R}_n^{-1}\mathbf{s}_n-\tilde{\mathcal{R}}_n^{-1}\tilde{\mathbf{s}}_n)=\mathcal{R}_n^{-1}(\tilde{\mathcal{R}}_n-\mathcal{R}_n)\tilde{\mathcal{R}}_n^{-1}\mathbf{s}_n+\tilde{\mathcal{R}}_n^{-1}(\mathbf{s}_n-\tilde{\mathbf{s}}_n).$$

Now, almost surely $\mathcal{R}_n^{-1},\tilde{\mathcal{R}}_n^{-1}=O(1)$ (if $\mathrm{E}(\mathcal{R}_n)$ is non-singular). Hence we only need to obtain a

bound for $\tilde{\mathcal{R}}_n - \mathcal{R}_n$ and $\mathbf{s}_n - \tilde{\mathbf{s}}_n$. We recall that

$$\tilde{\mathcal{R}}_n - \mathcal{R}_n = \frac{1}{n} \sum_{t=p_n+1} (\tilde{\mathbf{Y}}_t \tilde{\mathbf{Y}}_t' - \mathbf{Y}_t \mathbf{Y}_t'),$$

hence the terms differ where we replace the estimated $\tilde{\varepsilon}_t$ with the true ε_t, hence by using (14.23) and (14.24) we have almost surely

$$|\tilde{\mathcal{R}}_n - \mathcal{R}_n| = O(p_n Q(n) + \rho^{p_n}) \text{ and } |\tilde{\mathbf{s}}_n - \mathbf{s}_n| = O(p_n Q(n) + \rho^{p_n}).$$

Therefore by substituting the above into (14.26) we obtain

$$\big(\tilde{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_0\big) \quad = \quad \mathcal{R}_n^{-1}\big(\mathbf{s}_n - \mathcal{R}_n \boldsymbol{\varphi}_0\big) + O(p_n Q(n) + \rho^{p_n}). \tag{14.26}$$

Finally using straightforward algebra it can be shown that

$$\mathbf{s}_n - \mathcal{R}_n \boldsymbol{\varphi}_n = \frac{1}{n} \sum_{t=\max(p,q)}^{n} \varepsilon_t \mathbf{Y}_t.$$

By using Theorem 14.7.3, below, we have $\mathbf{s}_n - \mathcal{R}_n \boldsymbol{\varphi}_n = O((p+q)\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}})$. Substituting the above bound into (**??**), and noting that $O(Q(n))$ dominates $O(\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}})$ gives

$$\big\|\tilde{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_n\big\|_2 \quad = \quad O\left(p_n^3 \sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}} + \frac{p_n^4}{n} + p_n^2 \rho^{p_n}\right)$$

and the required result. □

14.7.1 Proof of Theorem 14.7.1 (A rate for $\|\hat{\mathbf{b}}_T - \mathbf{b}_T\|_2$)

We observe that

$$\hat{\mathbf{b}}_n - \mathbf{b}_n = R_n^{-1}\big(\hat{\mathbf{r}}_n - \hat{R}_n \mathbf{b}_n\big) + \big(\hat{R}_n^{-1} - R_n^{-1}\big)\big(\hat{\mathbf{r}}_n - \hat{R}_n \mathbf{b}_n\big)$$

where $\mathbf{b}$, R_n and $\mathbf{r}_n$ are deterministic, with $\mathbf{b}_n = (b_1 \ldots, b_{p_n})$, $(R_n)_{i,j} = \mathrm{E}(X_i X_j)$ and $(\mathbf{r}_n)_i = \mathrm{E}(X_0 X_{-i})$. Evaluating the Euclidean distance we have

$$\|\hat{\mathbf{b}}_n - \mathbf{b}_n\|_2 \leq \|R_n^{-1}\|_{spec} \big\|\hat{\mathbf{r}}_n - \hat{R}_n \mathbf{b}_n\big\|_2 + \|R_n^{-1}\|_{spec} \|\hat{R}_n^{-1}\|_{spec} \big\|\hat{R}_n - R_n\big\|_2 \big\|\hat{\mathbf{r}}_n - \hat{R}_n \mathbf{b}_n\big\|_2, \quad (14.27)$$

where we used that $\hat{R}_n^{-1} - \hat{R}_n^{-1} = \hat{R}_n^{-1}(R_n - \hat{R}_n)R_n^{-1}$ and the norm inequalities. Now by using Lemma 7.14.2 we have $\lambda_{min}(R_n^{-1}) > \delta/2$ for all T. Thus our aim is to obtain almost sure bounds for $\|\hat{\mathbf{r}}_n - \hat{R}_n \mathbf{b}_n\|_2$ and $\|\hat{R}_n - R_n\|_2$, which requires the lemma below.

Theorem 14.7.3 *Let us suppose that $\{X_t\}$ has an ARMA representation where the roots of the characteristic polynomials $\phi(z)$ and $\theta(z)$ lie are greater than $1 + \delta$. Then*

(i)

$$\frac{1}{n} \sum_{t=r+1}^{n} \varepsilon_t X_{t-r} = O\Big(\sqrt{\frac{(\log\log n)^{1+\gamma} \log n}{n}}\Big) \quad (14.28)$$

(ii)

$$\frac{1}{n} \sum_{t=\max(i,j)}^{n} X_{t-i} X_{t-j} = O\Big(\sqrt{\frac{(\log\log n)^{1+\gamma} \log n}{n}}\Big). \quad (14.29)$$

for any $\gamma > 0$.

PROOF. The result is proved in Chapter B.2. □

To obtain the bounds we first note that if the there wasn't an MA component in the ARMA process, in other words $\{X_t\}$ was an AR(p) process with $p_n \geq p$, then $\hat{\mathbf{r}}_n - \hat{R}_n \mathbf{b}_n = \frac{1}{n} \sum_{t=p_n+1}^{n} \varepsilon_t X_{t-r}$, which has a mean zero. However because an ARMA process has an AR(∞) representation and we are only estimating the first p_n parameters, there exists a 'bias' in $\hat{\mathbf{r}}_n - \hat{R}_n \mathbf{b}_n$. Therefore we obtain the decomposition

$$\begin{aligned} (\hat{\mathbf{r}}_n - \hat{R}_n \mathbf{b}_n)_r &= \frac{1}{n} \sum_{t=p_n+1}^{n} \big(X_t - \sum_{j=1}^{\infty} b_j X_{t-j}\big) X_{t-r} + \frac{1}{n} \sum_{t=p_n+1}^{n} \sum_{j=p_n+1}^{\infty} b_j X_{t-j} X_{t-r} \quad (14.30) \\ &= \underbrace{\frac{1}{n} \sum_{t=p_n+1}^{n} \varepsilon_t X_{t-r}}_{\text{stochastic term}} + \underbrace{\frac{1}{n} \sum_{t=p_n+1}^{n} \sum_{j=p_n+1}^{\infty} b_j X_{t-j} X_{t-r}}_{\text{bias}} \quad (14.31) \end{aligned}$$

Therefore we can bound the bias with

$$\left|(\hat{\mathbf{r}}_n - \hat{R}_n\mathbf{b}_n)_r - \frac{1}{n}\sum_{t=p_n+1}^{n}\varepsilon_t X_{t-r}\right| \leq K\rho^{p_n}\frac{1}{n}\sum_{t=1}^{n}|X_{t-r}|\sum_{j=1}^{\infty}\rho^j|X_{t-p_n-j}|. \tag{14.32}$$

Let $Y_t = \sum_{j=1}^{\infty}\rho^j|X_{t-j}$ and $S_{n,k,r} = \frac{1}{n}\sum_{t=1}^{n}|X_{t-r}|\sum_{j=1}^{\infty}\rho^j|X_{t-k-j}|$. We note that $\{Y_t\}$ and $\{X_t\}$ are ergodic sequences. By applying the ergodic theorm we can show that for a fixed k and r, $S_{n,k,r} \xrightarrow{\text{a.s.}} \mathrm{E}(X_{t-r}Y_{t-k})$. Hence $S_{n,k,r}$ are almost surely bounded sequences and

$$\rho^{p_n}\frac{1}{n}\sum_{t=1}^{n}|X_{t-r}|\sum_{j=1}^{\infty}\rho^j|X_{t-p_n-j}| = O(\rho^{p_n}).$$

Therefore almost surely we have

$$\|\hat{\mathbf{r}}_n - \hat{R}_n\mathbf{b}_n\|_2 = \|\frac{1}{n}\sum_{t=p_n+1}^{n}\varepsilon_t\mathbf{X}_{t-1}\|_2 + O(p_n\rho^{p_n}).$$

Now by using (14.28) we have

$$\|\hat{\mathbf{r}}_n - \hat{R}_n\mathbf{b}_n\|_2 = O\left(p_n\left\{\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}} + \rho^{p_n}\right\}\right). \tag{14.33}$$

This gives us a rate for $\hat{\mathbf{r}}_n - \hat{R}_n\mathbf{b}_n$. Next we consider $\hat{R}_n$. It is clear from the definition of $\hat{R}_n$ that almost surely we have

$$\begin{aligned}
(\hat{R}_n)_{i,j} - \mathrm{E}(X_iX_j) &= \frac{1}{n}\sum_{t=p_n+1}^{n}X_{t-i}X_{t-j} - \mathrm{E}(X_iX_j) \\
&= \frac{1}{n}\sum_{t=\min(i,j)}^{n}[X_{t-i}X_{t-j} - \mathrm{E}(X_iX_j)] - \frac{1}{n}\sum_{t=\min(i,j)}^{p_n}X_{t-i}X_{t-j} + \frac{\min(i,j)}{n}\mathrm{E}(X_iX_j) \\
&= \frac{1}{n}\sum_{t=\min(i,j)}^{T}[X_{t-i}X_{t-j} - \mathrm{E}(X_iX_j)] + O(\frac{p_n}{n}).
\end{aligned}$$

Now by using (14.29) we have almost surely

$$|(\hat{R}_n)_{i,j} - \mathrm{E}(X_iX_j)| = O(\frac{p_n}{n} + \sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}}).$$

Therefore we have almost surely

$$\|\hat{R}_n - R_n\|_2 = O\left(p_n^2\left\{\frac{p_n}{n} + \sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}}\right\}\right). \tag{14.34}$$

We note that by using (14.27), (14.33) and (14.34) we have

$$\|\hat{\mathbf{b}}_n - \mathbf{b}_n\|_2 \leq \|R_n^{-1}\|_{spec}\|\hat{R}_n^{-1}\|_{spec} O\left(p_n^2\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}} + \frac{p_n^2}{n} + p_n\rho^{p_n}\right).$$

As we mentioned previously, because the spectrum of X_t is bounded away from zero, $\lambda_{min}(R_n)$ is bounded away from zero for all T. Moreover, since $\lambda_{min}(\hat{R}_n) \geq \lambda_{min}(R_n) - \lambda_{max}(\hat{R}_n - R_n) \geq \lambda_{min}(R_n) - tr((\hat{R}_n - R_n)^2)$, which for a large enough n is bounded away from zero. Hence we obtain almost surely

$$\|\hat{\mathbf{b}}_n - \mathbf{b}_n\|_2 = O\left(p_n^2\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}} + \frac{p_n^3}{n} + p_n\rho^{p_n}\right), \tag{14.35}$$

thus proving Theorem 14.7.1 for any $\gamma > 0$.

14.8 Asymptotic properties of the GMLE

Let us suppose that $\{X_t\}$ satisfies the ARMA representation

$$X_t - \sum_{i=1}^{p}\phi_i^{(0)}X_{t-i} = \varepsilon_t + \sum_{j=1}^{q}\theta_j^{(0)}\varepsilon_{t-j}, \tag{14.36}$$

and $\boldsymbol{\theta}_0 = (\theta_1^{(0)}, \ldots, \theta_q^{(0)})$, $\boldsymbol{\phi}_0 = (\phi_1^{(0)}, \ldots, \phi_p^{(0)})$ and $\sigma_0^2 = \mathrm{var}(\varepsilon_t)$. In this section we consider the sampling properties of the GML estimator, defined in Section 9.2.1. We first recall the estimator. We use as an estimator of $(\underline{\theta}_0, \underline{\phi}_0)$, $\hat{\boldsymbol{\phi}}_n = (\hat{\underline{\theta}}_n, \hat{\underline{\phi}}_n, \hat{\boldsymbol{\sigma}}_n) = \arg\min_{(\underline{\theta},\underline{\phi})\in\Theta} L_n(\underline{\phi}, \underline{\theta}, \sigma)$, where

$$\frac{1}{n}L_n(\underline{\phi}, \underline{\theta}, \sigma) = \frac{1}{n}\sum_{t=1}^{n-1}\log r_{t+1}(\sigma, \underline{\phi}, \underline{\theta}) + \frac{1}{n}\sum_{t=1}^{n-1}\frac{(X_{t+1} - X_{t+1|t}^{(\phi,\theta)})^2}{r_{t+1}(\sigma, \underline{\phi}, \underline{\theta})}. \tag{14.37}$$

To show consistency and asymptotic normality we will use the following assumptions.

Assumption 14.8.1 *(i) X_t is both invertible and causal.*

(ii) The parameter space should be such that all $\phi(z)$ and $\theta(z)$ in the parameter space have roots whose absolute value is greater than $1+\delta$. $\phi_0(z)$ and $\theta_0(z)$ belong to this space.

Assumption 14.8.1 means for for some finite constant K and $\frac{1}{1+\delta} \leq \rho < 1$, we have $|\phi(z)^{-1}| \leq K\sum_{j=0}^{\infty}|\rho^j||z^j|$ and $|\phi(z)^{-1}| \leq K\sum_{j=0}^{\infty}|\rho^j||Z^j|$.

To prove the result, we require the following approximations of the GML. Let

$$\tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)} = \sum_{j=1}^{t} b_j(\underline{\phi},\underline{\theta})X_{t+1-j}. \tag{14.38}$$

This is an approximation of the one-step ahead predictor. Since the likelihood is constructed from the one-step ahead predictors, we can approximated the likelihood $\frac{1}{n}L_n(\underline{\phi},\underline{\theta},\sigma)$ with the above and define

$$\frac{1}{n}\tilde{\mathcal{L}}_n(\underline{\phi},\underline{\theta},\sigma) = \log\sigma^2 + \frac{1}{n\sigma^2}\sum_{t=1}^{T-1}(X_{t+1} - \tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)})^2. \tag{14.39}$$

We recall that $\tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)}$ was derived from $X_{t+1|t,\ldots}^{(\phi,\theta)}$ which is the one-step ahead predictor of X_{t+1} given $X_t, X_{t-1},\ldots$, this is

$$X_{t+1|t,\ldots}^{(\phi,\theta)} = \sum_{j=1}^{\infty} b_j(\underline{\phi},\underline{\theta})X_{t+1-j}. \tag{14.40}$$

Using the above we define a approximation of $\frac{1}{n}L_n(\underline{\phi},\underline{\theta},\sigma)$ which in practice cannot be obtained (since the infinite past of $\{X_t\}$ is not observed). Let us define the criterion

$$\frac{1}{n}\mathcal{L}_n(\underline{\phi},\underline{\theta},\sigma) = \log\sigma^2 + \frac{1}{n\sigma^2}\sum_{t=1}^{T-1}(X_{t+1} - X_{t+1|t,\ldots}^{(\phi,\theta)})^2. \tag{14.41}$$

In practice $\frac{1}{n}\mathcal{L}_n(\underline{\phi},\underline{\theta},\sigma)$ can not be evaluated, but it proves to be a convenient tool in obtaining the sampling properties of $\hat{\boldsymbol{\phi}}_n$. The main reason is because $\frac{1}{n}\mathcal{L}_n(\underline{\phi},\underline{\theta},\sigma)$ is a function of $\{X_t\}$ and $\{X_{t+1|t,\ldots}^{(\phi,\theta)} = \sum_{j=1}^{\infty} b_j(\underline{\phi},\underline{\theta})X_{t+1-j}\}$ both of these are ergodic (since the ARMA process is ergodic when its roots lie outside the unit circle and the roots of $\phi,\theta \in \Theta$ are such that they lie outside the unit circle). In contrast looking at $L_n(\underline{\phi},\underline{\theta},\sigma)$, which is comprised of $\{X_{t+1|t}\}$, which not an ergodic random variable because X_{t+1} is the best linear predictor of X_{t+1} given $X_t,\ldots,X_1$ (see the number of elements in the prediction changes with t). Using this approximation really simplifies the proof, though it is possible to prove the result without using these approximations.

First we obtain the result for the estimators $\hat{\varphi}_n^* = (\underline{\theta}_n^*, \underline{\phi}_n^*, \hat{\sigma}_n) = \arg\min_{(\underline{\theta},\underline{\phi})\in\Theta} \mathcal{L}_n(\underline{\phi}, \underline{\theta}, \sigma)$ and then show the same result can be applied to $\hat{\varphi}_n$.

Proposition 14.8.1 *Suppose $\{X_t\}$ is an ARMA process which satisfies (14.36), and Assumption 14.8.1 is satisfied. Let $X_{t+1|t}^{(\phi,\theta)}$, $\tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)}$ and $X_{t+1|t,\ldots}^{(\phi,\theta)}$ be the predictors defined in (**??**), (14.38) and (14.40), obtained using the parameters $\phi = \{\phi_j\}$ and $\theta = \{\theta_i\}$, where the roots the corresponding characteristic polynomial $\phi(z)$ and $\theta(z)$ have absolute value greater than $1+\delta$. Then*

$$\big|X_{t+1|t}^{(\phi,\theta)} - \tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)}\big| \leq \frac{\rho^t}{1-\rho}\sum_{i=1}^{t}\rho^i|X_i|, \tag{14.42}$$

$$\mathrm{E}(X_{t+1|t}^{(\phi,\theta)} - \tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)})^2 \leq K\rho^t, \tag{14.43}$$

$$\big|\tilde{X}_{t+1|t,\ldots}(1) - X_{t+1|t,\ldots}\big| = \big|\sum_{j=t+1}^{\infty} b_j(\phi,\theta)X_{t+1-j}\big| \leq K\rho^t\sum_{j=0}^{\infty}\rho^j|X_{-j}|, \tag{14.44}$$

$$\mathrm{E}(X_{t+1|t,\ldots}^{(\phi,\theta)} - \tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)})^2 \leq K\rho^t \tag{14.45}$$

and

$$|r_t(\sigma,\underline{\phi},\underline{\theta}) - \sigma^2| \leq K\rho^t \tag{14.46}$$

for any $1/(1+\delta) < \rho < 1$ and K is some finite constant.

PROOF. The proof follows closely the proof of Proposition 14.8.1. First we define a separate ARMA process $\{Y_t\}$, which is driven by the parameters θ and ϕ (recall that $\{X_t\}$ is drive by the parameters θ_0 and ϕ_0). That is Y_t satisfies $Y_t - \sum_{j=1}^{p}\phi_j Y_{t-j} = \varepsilon_t + \sum_{j=1}^{q}\theta_j\varepsilon_{t-j}$. Recalling that $X_{t+1|t}^{\phi,\theta}$ is the best linear predictor of X_{t+1} given $X_t,\ldots,X_1$ and the variances of $\{Y_t\}$ (noting that it is the process driven by θ and ϕ), we have

$$X_{t+1|t}^{\phi,\theta} = \sum_{j=1}^{t} b_j(\phi,\theta)X_{t+1-j} + \big(\sum_{j=t+1}^{\infty} b_j(\phi,\theta)\boldsymbol{r}_{t,j}'(\phi,\theta)\Sigma_t(\phi,\theta)^{-1}\big)\boldsymbol{X}_t, \tag{14.47}$$

where $\Sigma_t(\phi,\theta)_{s,t} = \mathrm{E}(Y_s Y_t)$, $(\boldsymbol{r}_{t,j})_i = \mathrm{E}(Y_{t-i}Y_{-j})$ and $\boldsymbol{X}_t' = (X_t,\ldots,X_1)$. Therefore

$$X_{t+1|t}^{\phi,\theta} - \tilde{X}_{t+1|t,\ldots} = \big(\sum_{j=t+1}^{\infty} b_j \boldsymbol{r}_{t,j}'\Sigma_t(\phi,\theta)^{-1}\big)\boldsymbol{X}_t.$$

Since the largest eigenvalue of $\Sigma_t(\phi,\theta)^{-1}$ is bounded (see Lemma 7.14.2) and $|(\boldsymbol{r}_{t,j})_i| = |\mathrm{E}(Y_{t-i}Y_{-j})| \leq K\rho^{|t-i+j|}$ we obtain the bound in (14.42). Taking expectations, we have

$$\mathrm{E}(X_{t+1|t}^{\phi,\theta} - \tilde{X}_{t+1|t,\ldots}^{\phi,\theta})^2 = \big(\sum_{j=t+1}^{\infty} b_j \boldsymbol{r}_{t,j}'\big)\Sigma_t(\phi,\theta)^{-1}\Sigma_t(\phi_0,\theta_0)\Sigma_t(\phi,\theta)^{-1}\big(\sum_{j=t+1}^{\infty} b_{t+j}\boldsymbol{r}_{t,j}\big).$$

Now by using the same arguments given in the proof of (7.23) we obtain (14.43).

To prove (14.45) we note that

$$\mathrm{E}(\tilde{X}_{t+1|t,\ldots}(1) - X_{t+1|t,\ldots})^2 = \mathrm{E}(\sum_{j=t+1}^{\infty} b_j(\phi,\theta)X_{t+1-j})^2 = \mathrm{E}(\sum_{j=1}^{\infty} b_{t+j}(\phi,\theta)X_{-j})^2,$$

now by using (4.23), we have $|b_{t+j}(\phi,\theta)| \leq K\rho^{t+j}$, for $\frac{1}{1+\delta} < \rho < 1$, and the bound in (14.44). Using this we have $\mathrm{E}(\tilde{X}_{t+1|t,\ldots}(1) - X_{t+1|t,\ldots})^2 \leq K\rho^t$, which proves the result. □

Using $\varepsilon_t = X_t - \sum_{j=1}^{\infty} b_j(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0)X_{t-j}$ and substituting this into $\mathcal{L}_n(\boldsymbol{\phi},\boldsymbol{\theta},\sigma)$ gives

$$\begin{aligned}
\frac{1}{n}\mathcal{L}_n(\boldsymbol{\phi},\boldsymbol{\theta},\sigma) &= \log\sigma^2 + \frac{1}{n\sigma^2}\big(X_t - \sum_{j=1}^{\infty} b_j(\underline{\phi},\underline{\theta})X_{t+1-j}\big)^2 \\
&= \frac{1}{n}\mathcal{L}_n(\underline{\phi},\underline{\theta},\sigma)\log\sigma^2 + \frac{1}{n\sigma^2}\sum_{t=1}^{T-1}\big\{\theta(B)^{-1}\phi(B)X_t\big\}\big\{\theta(B)^{-1}\phi(B)X_t\big\} \\
&= \log\sigma^2 + \frac{1}{n\sigma^2}\sum_{t=1}^{n}\varepsilon_t^2 + \frac{2}{n}\sum_{t=1}^{n}\varepsilon_t\big(\sum_{j=1}^{\infty} b_j(\boldsymbol{\phi},\boldsymbol{\theta})X_{t-j}\big) \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\big(\sum_{j=1}^{\infty}(b_j(\boldsymbol{\phi},\boldsymbol{\theta}) - b_j(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0))X_{t-j}\big)^2.
\end{aligned}$$

Remark 14.8.1 (Derivatives involving the Backshift operator) *Consider the transformation*

$$\frac{1}{1-\theta B}X_t = \sum_{j=0}^{\infty}\theta^j B^j X_t = \sum_{j=0}^{\infty}\theta^j X_{t-j}.$$

Suppose we want to differentiate the above with respect to θ, there are two ways this can be done. Either differentiate $\sum_{j=0}^{\infty}\theta^j X_{t-j}$ with respect to θ or differentiate $\frac{1}{1-\theta B}$ with respect to θ. In other

words

$$\frac{d}{d\theta}\frac{1}{1-\theta B}X_t = \frac{-B}{(1-\theta B)^2}X_t = \sum_{j=0}^{\infty} j\theta^{j-1}X_{t-j}.$$

Often it is easier to differentiate the operator. Suppose that $\theta(B) = 1 + \sum_{j=1}^{p}\theta_j B^j$ *and* $\phi(B) = 1 - \sum_{j=1}^{q}\phi_j B^j$*, then we have*

$$\frac{d}{d\theta_j}\frac{\phi(B)}{\theta(B)}X_t = -\frac{B^j\phi(B)}{\theta(B)^2}X_t = -\frac{\phi(B)}{\theta(B)^2}X_{t-j}$$
$$\frac{d}{d\phi_j}\frac{\phi(B)}{\theta(B)}X_t = -\frac{B^j}{\theta(B)^2}X_t = -\frac{1}{\theta(B)^2}X_{t-j}.$$

Moreover in the case of squares we have

$$\frac{d}{d\theta_j}(\frac{\phi(B)}{\theta(B)}X_t)^2 = -2(\frac{\phi(B)}{\theta(B)}X_t)(\frac{\phi(B)}{\theta(B)^2}X_{t-j}), \qquad \frac{d}{d\phi_j}(\frac{\phi(B)}{\theta(B)}X_t)^2 = -2(\frac{\phi(B)}{\theta(B)}X_t)(\frac{1}{\theta(B)^2}X_{t-j}).$$

Using the above we can easily evaluate the gradient of $\frac{1}{n}\mathcal{L}_n$

$$\begin{aligned}
\frac{1}{n}\nabla_{\theta_i}\mathcal{L}_n(\boldsymbol{\phi},\boldsymbol{\theta},\sigma) &= -\frac{2}{\sigma^2}\sum_{t=1}^{n}(\theta(B)^{-1}\phi(B)X_t)\frac{\phi(B)}{\theta(B)^2}X_{t-i} \\
\frac{1}{n}\nabla_{\phi_j}\mathcal{L}_n(\boldsymbol{\phi},\boldsymbol{\theta},\sigma) &= -\frac{2}{n\sigma^2}\sum_{t=1}^{n}(\theta(B)^{-1}\phi(B)X_t)\frac{1}{\theta(B)}X_{t-j} \\
\frac{1}{n}\nabla_{\sigma^2}\mathcal{L}_n(\boldsymbol{\phi},\boldsymbol{\theta},\sigma) &= \frac{1}{\sigma^2} - \frac{1}{n\sigma^4}\sum_{t=1}^{n}\big(X_t - \sum_{j=1}^{\infty}b_j(\boldsymbol{\phi},\boldsymbol{\theta})X_{t-j}\big)^2. \qquad (14.48)
\end{aligned}$$

Let $\nabla = (\nabla_{\phi_i}, \nabla_{\theta_j}, \nabla_{\sigma^2})$. We note that the second derivative $\nabla^2\mathcal{L}_n$ can be defined similarly.

Lemma 14.8.1 *Suppose Assumption 14.8.1 holds. Then*

$$\sup_{\underline{\phi},\underline{\theta}\in\Theta}\|\frac{1}{n}\nabla\mathcal{L}_n\|_2 \leq KS_n \qquad \sup_{\underline{\phi},\underline{\theta}\in\Theta}\|\frac{1}{n}\nabla^3\mathcal{L}_n\|_2 \leq KS_n \qquad (14.49)$$

for some constant K*,*

$$S_n = \frac{1}{n}\sum_{r_1,r_2=0}^{\max(p,q)}\sum_{t=1}^{n}Y_{t-r_1}Y_{t-r_2} \qquad (14.50)$$

where

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

for any $\frac{1}{(1+\delta)} < \rho < 1$.

PROOF. The proof follows from the the roots of $\phi(z)$ and $\theta(z)$ having absolute value greater than $1+\delta$. □

Define the expectation of the likelihood $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma)) = \mathrm{E}(\frac{1}{n}\mathcal{L}_n(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma))$. We observe

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma)) = \log \sigma^2 + \frac{\sigma_0^2}{\sigma_2} + \frac{1}{\sigma^2}\mathrm{E}(Z_t(\boldsymbol{\phi}, \boldsymbol{\theta})^2)$$

where

$$Z_t(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{j=1}^{\infty} (b_j(\boldsymbol{\phi}, \boldsymbol{\theta}) - b_j(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0)) X_{t-j}$$

Lemma 14.8.2 *Suppose that Assumption 14.8.1 are satisfied. Then for all* $\underline{\theta}, \underline{\phi}, \theta \in \Theta$ *we have*

(i) $\frac{1}{n}\nabla^i \mathcal{L}_n(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma)) \stackrel{a.s.}{\to} \nabla^i \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma))$ *for* $i = 0, 1, 2, 3$.

(ii) Let S_n *defined in (14.50), then* $S_n \stackrel{a.s.}{\to} \mathrm{E}(\sum_{r_1, r_2=0}^{\max(p,q)} \sum_{t=1}^{n} Y_{t-r_1} Y_{t-r_2})$.

PROOF. Noting that the ARMA process $\{X_t\}$ are ergodic random variables, then $\{Z_t(\boldsymbol{\phi}, \boldsymbol{\theta})\}$ and $\{Y_t\}$ are ergodic random variables, the result follows immediately from the Ergodic theorem.

We use these results in the proofs below.

Theorem 14.8.1 *Suppose that Assumption 14.8.1 is satisfied. Let* $(\hat{\underline{\theta}}_n^*, \hat{\underline{\phi}}_n^*, \hat{\sigma}_n^*) = \arg\min \mathcal{L}_n(\underline{\theta}, \underline{\phi}, \sigma)$ *(noting the practice that this cannot be evaluated). Then we have*

(i) $(\hat{\underline{\theta}}_n^*, \hat{\underline{\phi}}_n^*, \hat{\sigma}_n^*) \stackrel{a.s.}{\to} (\underline{\theta}_0, \underline{\phi}_0, \sigma_0)$.

(ii) $\sqrt{n}(\hat{\underline{\theta}}_n^* - \underline{\theta}_0, \hat{\underline{\phi}}_n^* - \underline{\theta}_0) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_0^2 \Lambda^{-1})$, *where*

$$\Lambda = \begin{pmatrix} \mathrm{E}(U_t U_t') & \mathrm{E}(V_t U_t') \\ \mathrm{E}(U_t V_t') & \mathrm{E}(V_t V_t') \end{pmatrix}$$

and $\{U_t\}$ *and* $\{V_t\}$ *are autoregressive processes which satisfy* $\phi_0(B)U_t = \varepsilon_t$ *and* $\theta_0(B)V_t = \varepsilon_t$.

PROOF. We prove the result in two stages below. □

PROOF of Theorem 14.8.1(i) We will first prove Theorem 14.8.1(i). Noting the results in Section 14.3, to prove consistency we recall that we must show (a) the $(\underline{\phi}_0, \underline{\theta}_0, \sigma_0)$ is the unique minimum of $\mathcal{L}(\cdot)$ (b) pointwise convergence $\frac{1}{T}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma)) \stackrel{\text{a.s.}}{\to} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma))$ and (b) stochastic equicontinuity (as defined in Definition 14.3.2). To show that $(\underline{\phi}_0, \underline{\theta}_0, \sigma_0)$ is the minimum we note that

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma)) - \mathcal{L}(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \sigma_0)) = \log(\frac{\sigma^2}{\sigma_0^2}) + \frac{\sigma^2}{\sigma_0^2} - 1 + \mathrm{E}(Z_t(\boldsymbol{\phi}, \boldsymbol{\theta})^2).$$

Since for all positive x, $\log x + x - 1$ is a positive function and $\mathrm{E}(Z_t(\boldsymbol{\phi}, \boldsymbol{\theta})^2) = \mathrm{E}(\sum_{j=1}^{\infty}(b_j(\boldsymbol{\phi}, \boldsymbol{\theta}) - b_j(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0))X_{t-j})^2$ is positive and zero at $(\underline{\phi}_0, \underline{\theta}_0, \sigma_0)$ it is clear that $\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \sigma_0$ is the minimum of $\mathcal{L}$. We will assume for now it is the unique minimum. Pointwise convergence is an immediate consequence of Lemma 14.8.2(i). To show stochastic equicontinuity we note that for any $\boldsymbol{\varphi}_1 = (\boldsymbol{\phi}_1, \boldsymbol{\theta}_1, \sigma_1)$ and $\boldsymbol{\varphi}_2 = (\boldsymbol{\phi}_2, \boldsymbol{\theta}_2, \sigma_2)$ we have by the mean value theorem

$$\mathcal{L}_n(\boldsymbol{\phi}_1, \boldsymbol{\theta}_1, \sigma_1) - \mathcal{L}_n(\boldsymbol{\phi}_2, \boldsymbol{\theta}_2, \sigma_2)) = (\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2)\nabla\mathcal{L}_n(\bar{\boldsymbol{\phi}}, \bar{\boldsymbol{\theta}}, \bar{\sigma}).$$

Now by using (14.49) we have

$$\mathcal{L}_n(\boldsymbol{\phi}_1, \boldsymbol{\theta}_1, \sigma_1) - \mathcal{L}_n(\boldsymbol{\phi}_2, \boldsymbol{\theta}_2, \sigma_2)) \leq S_T\|(\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2), (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2), (\sigma_1 - \sigma_2)\|_2.$$

By using Lemma 14.8.2(ii) we have $S_n \stackrel{\text{a.s.}}{\to} \mathrm{E}(\sum_{r_1,r_2=0}^{\max(p,q)}\sum_{t=1}^{n} Y_{t-r_1}Y_{t-r_2})$, hence $\{S_n\}$ is almost surely bounded. This implies that $\mathcal{L}_n$ is equicontinuous. Since we have shown pointwise convergence and equicontinuity of $\mathcal{L}_n$, by using Corollary 14.3.1, we almost sure convergence of the estimator. Thu proving (i). □

PROOF of Theorem 14.8.1(ii) We now prove Theorem 14.8.1(i) using the Martingale central limit theorem (see Billingsley (1995) and Hall and Heyde (1980)) in conjunction with the Cramer-Wold device (see Theorem 9.1.1).

Using the mean value theorem we have

$$\left(\hat{\boldsymbol{\varphi}}_n^* - \boldsymbol{\varphi}_0\right) = \nabla^2\mathcal{L}_n^*(\bar{\boldsymbol{\varphi}}_n)^{-1}\nabla\mathcal{L}_n^*(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \sigma_0)$$

where $\hat{\boldsymbol{\varphi}}_n^* = (\hat{\boldsymbol{\phi}}_n^*, \hat{\boldsymbol{\theta}}_n^*, \hat{\sigma}_n^*)$, $\boldsymbol{\varphi}_0 = (\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \sigma_0)$ and $\bar{\boldsymbol{\varphi}}_n = \bar{\boldsymbol{\phi}}, \bar{\boldsymbol{\theta}}, \bar{\sigma}$ lies between $\hat{\boldsymbol{\varphi}}_n^*$ and $\boldsymbol{\varphi}_0$.

Using the same techniques given in Theorem 14.8.1(i) and Lemma 14.8.2 we have pointwise convergence and equicontinuity of $\nabla^2\mathcal{L}_n$. This means that $\nabla^2\mathcal{L}_n(\bar{\varphi}_n) \stackrel{\text{a.s.}}{\to} \mathrm{E}(\nabla^2\mathcal{L}_n(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\sigma_0)) = \frac{1}{\sigma^2}\Lambda$ (since by definition of $\bar{\varphi}_n$ $\bar{\boldsymbol{\varphi}}_n \stackrel{\text{a.s.}}{\to} \boldsymbol{\varphi}_0$). Therefore by applying Slutsky's theorem (since Λ is nonsingular) we have

$$\nabla^2\mathcal{L}_n(\bar{\boldsymbol{\varphi}}_n)^{-1} \stackrel{\text{a.s.}}{\to} \sigma^2\Lambda^{-1}. \tag{14.51}$$

Now we show that $\nabla\mathcal{L}_n(\boldsymbol{\varphi}_0)$ is asymptotically normal. By using (14.48) and replacing $X_{t-i} = \phi_0(B)^{-1}\theta_0(B)\varepsilon_{t-i}$ we have

$$\begin{aligned}
\frac{1}{n}\nabla_{\theta_i}\mathcal{L}_n(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\sigma_0) &= \frac{2}{\sigma^2 n}\sum_{t=1}^{n}\varepsilon_t\frac{(-1)}{\theta_0(B)}\varepsilon_{t-i} = \frac{-2}{\sigma^2 n}\sum_{t=1}^{n}\varepsilon_t V_{t-i} \quad i=1,\ldots,q \\
\frac{1}{n}\nabla_{\phi_j}\mathcal{L}_n(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\sigma_0) &= \frac{2}{\sigma^2 n}\sum_{t=1}^{n}\varepsilon_t\frac{1}{\phi_0(B)}\varepsilon_{t-j} = \frac{2}{\sigma^2 n}\sum_{t=1}^{T}\varepsilon_t U_{t-j} \quad j=1,\ldots,p \\
\frac{1}{n}\nabla_{\sigma^2}L_n(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\sigma_0) &= \frac{1}{\sigma^2} - \frac{1}{\sigma^4 n}\sum_{t=1}^{T}\varepsilon^2 = \frac{1}{\sigma^4 n}\sum_{t=1}^{T}(\sigma^2-\varepsilon^2),
\end{aligned}$$

where $U_t = \frac{1}{\phi_0(B)}\varepsilon_t$ and $V_t = \frac{1}{\theta_0(B)}\varepsilon_t$. We observe that $\frac{1}{n}\nabla\mathcal{L}_n$ is the sum of vector martingale differences. If $\mathrm{E}(\varepsilon_t^4) < \infty$, it is clear that $\mathrm{E}((\varepsilon_t U_{t-j})^4) = \mathrm{E}((\varepsilon_t^4)\mathrm{E}(U_{t-j})^4) < \infty$, $\mathrm{E}((\varepsilon_t V_{t-i})^4) = \mathrm{E}((\varepsilon_t^4)\mathrm{E}(V_{t-i})^4) < \infty$ and $\mathrm{E}((\sigma^2-\varepsilon_t^2)^2) < \infty$. Hence Lindeberg's condition is satisfied (see the proof given in Section 9.1.6, for why this is true). Hence we have

$$\sqrt{n}\nabla\mathcal{L}_n(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\sigma_0) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0,\Lambda).$$

Now by using the above and (14.51) we have

$$\sqrt{n}\big(\hat{\boldsymbol{\varphi}}_n^* - \boldsymbol{\varphi}_0\big) = \sqrt{n}\nabla^2\mathcal{L}_n(\bar{\boldsymbol{\varphi}}_n)^{-1}\nabla\mathcal{L}_n(\boldsymbol{\varphi}_0) \Rightarrow \quad \sqrt{n}\big(\hat{\boldsymbol{\varphi}}_n^* - \boldsymbol{\varphi}_0\big) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0,\sigma^4\Lambda^{-1}).$$

Thus we obtain the required result. □

The above result proves consistency and asymptotically normality of $(\hat{\underline{\theta}}_n^*, \hat{\underline{\phi}}_n^*, \hat{\sigma}_n^*)$, which is based on $\mathcal{L}_n(\underline{\theta},\underline{\phi},\sigma)$, which in practice is impossible to evaluate. However we will show below that the gaussian likelihood, $L_n(\underline{\theta},\underline{\phi},\sigma)$ and is derivatives are sufficiently close to $\mathcal{L}_n(\underline{\theta},\underline{\phi},\sigma)$ such that the estimators $(\hat{\underline{\theta}}_n^*, \hat{\underline{\phi}}_n^*, \hat{\sigma}_n^*)$ and the GMLE, $(\hat{\underline{\theta}}_n, \hat{\underline{\phi}}_n, \hat{\sigma}_n) = \arg\min L_n(\underline{\theta},\underline{\phi},\sigma)$ are asymptotically equivalent. We use Lemma 14.8.1 to prove the below result.

Proposition 14.8.2 *Suppose that Assumption 14.8.1 hold and* $L_n(\underline{\theta},\underline{\phi},\sigma)$, $\tilde{\mathcal{L}}_n(\underline{\theta},\underline{\phi},\sigma)$ *and* $\mathcal{L}_n(\underline{\theta},\underline{\phi},\sigma)$ *are defined as in (14.37), (14.39) and (14.41) respectively. Then we have for all* $(\underline{\theta},\underline{\phi}) \in Theta$ *we have almost surely*

$$\sup_{(\boldsymbol{\phi},\boldsymbol{\theta},\sigma)} \frac{1}{n}|\nabla^{(k)}\tilde{\mathcal{L}}(\boldsymbol{\phi},\boldsymbol{\theta},\sigma) - \nabla^k L_n(\boldsymbol{\phi},\boldsymbol{\theta},\sigma)| = O(\frac{1}{n}) \quad \sup_{(\boldsymbol{\phi},\boldsymbol{\theta},\sigma)} \frac{1}{n}|\tilde{\mathcal{L}}_n(\boldsymbol{\phi},\boldsymbol{\theta},\sigma) - \mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta},\sigma)| = O(\frac{1}{n}),$$

for $k = 0, 1, 2, 3$.

PROOF. The proof of the result follows from (14.42) and (14.44). We show that result for $\sup_{(\boldsymbol{\phi},\boldsymbol{\theta},\sigma)} \frac{1}{n}|\tilde{\mathcal{L}}(\boldsymbol{\phi},\boldsymbol{\theta},\sigma) - L_n(\boldsymbol{\phi},\boldsymbol{\theta},\sigma)|$, a similar proof can be used for the rest of the result.

Let us consider the difference

$$\mathcal{L}_n(\underline{\phi},\underline{\theta}) - L_n(\underline{\phi},\underline{\theta}) = \frac{1}{n}(I_n + II_n + III_n),$$

where

$$\begin{aligned} I_n &= \sum_{t=1}^{n-1} \big\{r_t(\underline{\phi},\underline{\theta},\sigma) - \sigma^2\big\}, \quad II_n = \sum_{t=1}^{n-1} \frac{1}{r_t(\underline{\phi},\underline{\theta},\sigma)}(X_{t+1}^{(\phi,\theta)} - X_{t+1|t}^{(\phi,\theta)})^2 \\ III_n &= \sum_{t=1}^{n-1} \frac{1}{\sigma^2}\big\{2X_{t+1}(X_{t+1|t}^{(\phi,\theta)} - \tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)}) + ((X_{t+1|t}^{(\phi,\theta)})^2 - (\tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)})^2)\big\}. \end{aligned}$$

Now we recall from Proposition 14.8.1 that

$$\big|X_{t+1|t}^{(\phi,\theta)} - \tilde{X}_{t+1|t,\ldots}^{(\phi,\theta)}\big| \leq K \cdot V_t \frac{\rho^t}{(1-\rho)}$$

where $V_t = \sum_{i=1}^t \rho^i |X_i|$. Hence since $\mathrm{E}(X_t^2) < \infty$ and $\mathrm{E}(V_t^2) < \infty$ we have that $\sup_n \mathrm{E}|I_n| < \infty$, $\sup_n \mathrm{E}|II_n| < \infty$ and $\sup_n \mathrm{E}|III_n| < \infty$. Hence the sequence $\{|I_n + II_n + III_n|\}_n$ is almost surely bounded. This means that almost surely

$$\sup_{\underline{\phi},\underline{\theta},\sigma} \big|\mathcal{L}_n(\underline{\phi},\underline{\theta}) - L_n(\underline{\phi},\underline{\theta})\big| = O(\frac{1}{n}).$$

Thus giving the required result. □

Now by using the above proposition the result below immediately follows.

Theorem 14.8.2 *Let* $(\hat{\underline{\theta}},\hat{\underline{\phi}}) = \arg\min L_T(\underline{\theta},\underline{\phi},\sigma)$ *and* $(\tilde{\underline{\theta}},\hat{\underline{\phi}}) = \arg\min \tilde{L}_T(\underline{\theta},\underline{\phi},\sigma)$

(i) $(\hat{\underline{\theta}}, \hat{\underline{\phi}}) \overset{a.s.}{\to} (\underline{\theta}_0, \underline{\phi}_0)$ *and* $(\tilde{\underline{\theta}}, \tilde{\underline{\phi}}) \overset{a.s.}{\to} (\underline{\theta}_0, \underline{\phi}_0)$.

(ii) $\sqrt{T}(\hat{\underline{\theta}}_T - \underline{\theta}_0, \hat{\underline{\phi}}_T - \underline{\theta}_0) \overset{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_0^4 \Lambda^{-1})$

and $\sqrt{T}(\tilde{\underline{\theta}}_T - \underline{\theta}_0, \tilde{\underline{\phi}}_T - \underline{\theta}_0) \overset{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_0^4 \Lambda^{-1})$.

PROOF. The proof follows immediately from Proposition 14.8.1. □

Chapter 15

Residual Bootstrap for estimation in autoregressive processes

In Chapter **??** we consider the asymptotic sampling properties of the several estimators including the least squares estimator of the autoregressive parameters and the gaussian maximum likelihood estimator used to estimate the parameters of an ARMA process. The asymptotic distributions are often used for statistical testing and constructing confidence intervals. However the results are asymptotic, and only hold (approximately), when the sample size is relatively large. When the sample size is smaller, the normal approximation is not valid and better approximations are sought. Even in the case where we are willing to use the asymptotic distribution, often we need to obtain expressions for the variance or bias. Sometimes this may not be possible or only possible with a excessive effort. The Bootstrap is a power tool which allows one to approximate certain characteristics. To quote from Wikipedia 'Bootstrap is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution'. Bootstrap essentially samples from the sample. Each subsample is treated like a new sample from a population. Using these 'new' multiple realisations one can obtain approximations for CIs and variance estimates for the parameter estimates. Of course in reality we do not have multiple-realisations, we are sampling from the sample. Thus we are not gaining more as we subsample more. But we do gain some insight into the finite sample distribution. In this chapter we will details the residual bootstrap method, and then show that the asymptotically the bootstrap distribution coincides with asymptotic distribution.

The residual bootstrap method was first proposed by J. P. Kreiss (**?** is a very nice review paper

on the subject), (see also **?**, where an extension to AR(∞) processes is also given here). One of the first theoretical papers on the bootstrap is **?**. There are several other boostrapping methods for time series, these include bootstrapping the periodogram, block bootstrap, bootstrapping the Kalman filter (**?**, **?** and Shumway and Stoffer (2006)). These methods have not only been used for variance estimation but also determining orders etc. At this point it is worth mentioning methods Frequency domain approaches are considered in Dahlhaus and Janas (1996) and Franke and Härdle (1992) (a review of subsampling methods can be found in **?**).

15.1 The residual bootstrap

Suppose that the time series $\{X_t\}$ satisfies the stationary, causal AR process

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

where $\{\varepsilon_t\}$ are iid random variables with mean zero and variance one and the roots of the characteristic polynomial have absolute value greater than $(1+\delta)$. We will suppose that the order p is known.

The residual bootstrap for autoregressive processes

(i) Let

$$\hat{\Gamma}_p = \frac{1}{n}\sum_{t=p+1}^{n} \mathbf{X}_{t-1}\mathbf{X}_{t-1}' \text{ and } \hat{\gamma}_p = \frac{1}{n}\sum_{t=p+1}^{n} \mathbf{X}_{t-1}X_t, \tag{15.1}$$

where $\mathbf{X}_t' = (X_t, \ldots, X_{t-p+1})$. We use $\underline{\hat{\phi}}_n = (\hat{\phi}_1, \ldots, \hat{\phi}_p)' == \hat{\Gamma}_p^{-1}\hat{\gamma}_p$ as an estimator of $\underline{\phi} = (\phi_1, \ldots, \phi_p)$.

(ii) We create the bootstrap sample by first estimating the residuals $\{\varepsilon_t\}$ and sampling from the residuals. Let

$$\hat{\varepsilon}_t = X_t - \sum_{j=1}^{p} \phi_j X_{t-j}.$$

(iii) Now create the empirical distribution function based on $\hat{\varepsilon}_t$. Let

$$\hat{F}_n(x) = \frac{1}{n-p} \sum_{t=p+1}^{n} I_{(-\infty,\hat{\varepsilon}_t]}(x).$$

we notice that sampling from the distribution $\hat{F}_n(x)$, means observing $\hat{\varepsilon}_t$ with probability $(n-p)^{-1}$.

(iv) Sample independently from the distribution $\hat{F}_n(x)$ n times. Label this sample as $\{\varepsilon_k^+\}$.

(v) Let $X_k = \varepsilon_k^+$ for $1 \leq k \leq p$ and

$$X_k^+ = \sum_{j=1}^{p} \phi_j X_{k-j}^+ + \varepsilon_k, \quad p < k \leq n.$$

(vi) We call $\{X_k^+\}$. Repeating step (vi,v) N times gives us N bootstrap samples. To distinguish each sample we can label each bootstrap sample as $(\{(X_k^+)^{(i)}\}; i = p+1, \ldots, n)$.

(vii) For each bootstrap sample we can construct a bootstrap matrix, vector and estimator $(\Gamma_p^+)^{(i)}$, $(\gamma_p^+)^{(i)}$ and $(\hat{\underline{\phi}}_n^+)^{(i)} = ((\Gamma_p^+)^{(i)})^{-1}(\gamma_p^+)^{(i)}$.

(viii) Using $(\hat{\underline{\phi}}_n^+)^{(i)}$ we can estimate the variance of $\hat{\underline{\phi}}_n - \underline{\phi}$ with $\frac{1}{n}\sum_{j=1}^{n}((\hat{\underline{\phi}}_n^+)^{(i)} - \hat{\underline{\phi}}_n)$ and the distribution function of $\hat{\underline{\phi}}_n - \underline{\phi}$.

15.2 The sampling properties of the residual bootstrap estimator

In this section we show that the distribution of $\sqrt{n}(\hat{\underline{\phi}}_n^+ - \hat{\underline{\phi}}_n)$ and $\sqrt{n}(\hat{\underline{\phi}}_n - \underline{\phi})$ asymptotically coincide. This means that using the bootstrap distribution is no worse than using the asymptotic normal approximation. However it does not say the bootstrap distribution better approximates the finite sample distribution of $(\hat{\underline{\phi}}_n - \underline{\phi})$, to show this one would have to use Edgeworth expansion methods.

In order to show that the distribution of the bootstrap sample $\sqrt{n}(\hat{\underline{\phi}}_n^+ - \hat{\underline{\phi}}_n)$ asymptotically coincides with the asymptotic distribution of $\sqrt{n}(\hat{\underline{\phi}}_n - \underline{\phi})$, we will show convergence of the distributions

under the following distance

$$d_p(H,G) = \inf_{X\sim H, Y\sim G}\{\mathrm{E}(X-Y)^p\}^{1/p},$$

where $p > 1$. Roughly speaking, if $d_p(F_n, G_n) \to 0$, then the limiting distributions of F_n and G_n are the same (see **?**). The case that $p = 2$ is the most commonly used p, and for $p = 2$, this is called Mallows distance. The Mallows distance between the distribution H and G is defined as

$$d_2(H,G) = \inf_{X\sim H, Y\sim G}\{\mathrm{E}(X-Y)^2\}^{1/2},$$

we will use the Mallow distance to prove the results below. It is worth mentioning that the distance is zero when $H = G$ are the same (as a distance should be). To see this, set the joint distribution between X and Y to be $F(x,y) == G(x)$ when $y = x$ and zero otherwise, then it clear that $d_2(H,G) = 0$. To reduce notation rather than specify the distributions, F and G, we let $d_p(X,Y) = d_p(H,G)$, where the random variables X and Y have the marginal distributions H and G, respectively. We mention that distance d_p satisfies the triangle inequality.

The main application of showing that $d_p(F_n, G_n) \to 0$ is stated in the following lemma, which is a version of Lemma 8.3, **?**.

Lemma 15.2.1 *Let α, α_n be two probability measures then $d_p(\alpha_n, \alpha) \to 0$ if and only if*

$$\mathrm{E}_{\alpha_n}(|X|^p) = \int |x|^p \alpha_n(dx) \to \mathrm{E}_{\alpha}(|X|^p) = \int |x|^p \alpha(dx) \quad n \to \infty.$$

and the distribution α_n converges weakly to the distribution α.

Our aim is to show that

$$d_2\big(\sqrt{n}(\hat{\underline{\phi}}_n^+ - \hat{\underline{\phi}}_n), \sqrt{n}(\hat{\underline{\phi}}_n - \underline{\phi})\big) \to 0,$$

which implies that their distributions asymptotically coincide. To do this we use

$$\begin{aligned}(\sqrt{n}(\hat{\underline{\phi}}_n - \underline{\phi}) &= \sqrt{n}\hat{\Gamma}_p^{-1}(\hat{\gamma}_p - \hat{\Gamma}_p\underline{\phi}) \\ (\sqrt{n}(\hat{\underline{\phi}}_n^+ - \hat{\underline{\phi}}) &= \sqrt{n}(\Gamma_p^+)^{-1}(\gamma_p^+ - \Gamma_p^+\hat{\underline{\phi}}_n).\end{aligned}$$

Studying how $\hat{\Gamma}_p$, $\hat{\gamma}_p$, Γ_p^+ and γ_p^+ are constructed, we see as a starting point we need to show

$$d_2(X_t^+, X_t) \to 0 \quad t, n \to \infty, \quad d_2(Z_t^+, Z_t) \to 0 \quad n \to \infty.$$

We start by showing that $d_2(Z_t^+, Z_t) \to 0$

Lemma 15.2.2 *Suppose ε_t^+ is the bootstrap residuals and ε_t are the true residuals. Define the discrete random variable $J = \{p+1, \ldots, n\}$ and let $P(J = k) = \frac{1}{n-p}$. Then*

$$\mathrm{E}\big((\hat{\varepsilon}_J - \varepsilon_J)^2 | X_1, \ldots, X_n\big) = O_p(\frac{1}{n}) \tag{15.2}$$

and

$$d_2(\hat{F}_n, F) \quad \leq \quad d_2(\hat{F}_n, F_n) + d_2(F_n, F) \to 0 \quad as \ \to \infty, \tag{15.3}$$

where $F_n = \frac{1}{n-1}\sum_{t=p+1}^n I_{(-\infty,\varepsilon_t)}(x)$, $\hat{F}_n(x) = \frac{1}{n-p}\sum_{t=p+1}^n I_{(-\infty,\hat{\varepsilon}_t]}(x)$ are the empirical distribution function based on the residuals $\{\varepsilon_t\}_p^n$ and estimated residuals $\{\hat{\varepsilon}_t\}_p^n$, and F is the distribution function of the residual ε_t.

PROOF. We first show (15.2). From the definition of $\hat{\varepsilon}_J^+$ and ε_J we have

$$\begin{aligned}
\mathrm{E}(|\hat{\varepsilon}_J - \varepsilon_J|^2 | X_1, \ldots, X_n) &= \frac{1}{n-p}\sum_{t=p+1}^n (\hat{\varepsilon}_t - \varepsilon_t)^2 \\
&= \frac{1}{n-p}\sum_{t=p+1}^n (\sum_{j=1}^p [\hat{\phi}_j - \phi_j] X_{t-j})^2 \\
&= \sum_{j_1, j_2=1}^p [\hat{\phi}_{j_1} - \phi_{j_1}][\hat{\phi}_{j_2} - \phi_{j_2}] \frac{1}{n-p}\sum_{t=p+1}^n X_{t-j_1} X_{t-j_2}.
\end{aligned}$$

Now by using (14.14) we have $\sup_{1\leq j\leq p} |\hat{\phi}_j - \phi_j| = O_p(n^{-1/2})$, therefore we have $\mathrm{E}|\hat{\varepsilon}_J - \varepsilon_J|^2 = O_p(n^{-1/2})$.

We now prove (15.3). We first note by the triangle inequality we have

$$d_2(F, F_n) \leq d_2(F, F_n) + d_2(\hat{F}_n, F_n).$$

By using Lemma 8.4, **?**, we have that $d_2(F_n, F) \to 0$. Therefore we need to show that $d_2(\hat{F}_n, F_n) \to 0$. It is clear by definition that $d_2(\hat{F}_n, F_n) = d_2(\varepsilon_t^+, \tilde{\varepsilon}_t)$, where ε_t^+ is sampled from $\hat{F}_n = \frac{1}{n-1}\sum_{t=p+1}^n I_{(-\infty,\hat{\varepsilon}_t)}(x)$

and $\tilde{\varepsilon}_t$ is sampled from $F_n = \frac{1}{n-1}\sum_{t=p+1}^{n} I_{(-\infty,\varepsilon_t)}(x)$. Hence, $\tilde{\varepsilon}_t$ $\tilde{\varepsilon}_t^+$ have the same distribution as ε_J and $\hat{\varepsilon}_J$. We now evaluate $d_2(\varepsilon_t^+, \tilde{\varepsilon}_t)$. To evaluate $d_2(\varepsilon_t^+, \tilde{\varepsilon}_t) = \inf_{\varepsilon_t^+ \sim \hat{F}_n, \tilde{\varepsilon}_t \sim F_n} \mathrm{E}|\varepsilon_t^+ - \tilde{\varepsilon}_t|$ we need that the marginal distributions of $(\varepsilon_t^+, \tilde{\varepsilon}_t)$ are $\hat{F}_n$ and F_n, but the infimum is over all joint distributions. It is best to choose a joint distribution which is highly dependent (because this minimises the distance between the two random variables). An ideal candidate is to suppose that $\varepsilon_t^+ = \hat{\varepsilon}_J$ and $\tilde{\varepsilon}_t = \varepsilon_J$, since these have the marginals $\hat{F}_n$ and F_n respectively. Therefore

$$d_2(\hat{F}_n, F_n)^2 = \inf_{\varepsilon_t^+ \sim \hat{F}_n, \tilde{\varepsilon}_t \sim F_n} \mathrm{E}|\varepsilon_t^+ - \tilde{\varepsilon}_t|^2 \leq \mathrm{E}\big((\hat{\varepsilon}_J - \varepsilon_J)^2 | X_1, \ldots, X_n\big) = O_p(\frac{1}{n}),$$

where the above rate comes from (15.2). This means that $d_2(\hat{F}_n, F_n) \xrightarrow{\mathcal{P}} 0$, hence we obtain (15.3). □

Corollary 15.2.1 *Suppose ε_t^+ is the bootstrapped residual. Then we have*

$$\mathrm{E}_{\hat{F}_n}((\varepsilon_t^+)^2 | X_1, \ldots, X_n) \xrightarrow{\mathcal{P}} \mathrm{E}_F(\varepsilon_t^2)$$

PROOF. The proof follows from Lemma 15.2.1 and Lemma 15.2.2. □

We recall that since X_t is a causal autoregressive process, there exists some coefficients $\{a_j\}$ such that

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j},$$

where $a_j = a_j(\underline{\phi}) = [A(\phi)^j]_{1,1} = [A^j]_{1,1}$ (see Lemma 4.5.1). Similarly using the estimated parameters $\hat{\underline{\phi}}_n$ we can write X_t^+ as

$$X_t^+ = \sum_{j=0}^{t} a_j(\hat{\underline{\phi}}_n) \varepsilon_{t-j}^+,$$

where $a_j(\hat{\underline{\phi}}_n) = [A(\hat{\underline{\phi}}_n)^j]_{1,1}$. We now show that $d_2(X_t^+, X_t) \to 0$ as $n \to \infty$ and $t \to \infty$.

Lemma 15.2.3 *Let $J_{p+1}, \ldots, J_n$ be independent samples from $\{n-p+1, \ldots, n\}$ with $P(J_i = k) =$*

$\frac{1}{n-p}$. *Define*

$$Y_t^+ = \sum_{j=p+1}^{t} a_j(\hat{\underline{\phi}}_n)\varepsilon^+_{J_{t-j}}, \quad \tilde{Y}_t^+ = \sum_{j=p+1}^{t} a_j(\hat{\underline{\phi}}_n)\varepsilon^+_{J_{t-j}}, \quad \tilde{Y}_t = \sum_{j=p+1}^{t} a_j\varepsilon_{J_{t-j}}, \quad Y_t = \tilde{Y}_t + \sum_{j=t+p+1}^{\infty} a_j\varepsilon_{t-j},$$

where ε_{J_j} *is a sample from* $\{\varepsilon_{p+1},\ldots,\varepsilon_n\}$ *and* $\hat{\varepsilon}_J$ *is a sample from* $\{\hat{\varepsilon}_{p+1},\ldots,\hat{\varepsilon}_n\}$. *Then we have*

$$\mathrm{E}\big((Y_t^+ - \tilde{Y}_t^+)^2|X_1,\ldots,X_n\big) = O_p(\frac{1}{n}), \quad d_2(Y_t^+,\tilde{Y}_t^+)\to 0 \quad n\to\infty, \tag{15.4}$$

$$\mathrm{E}\big((\tilde{Y}_t^+ - \tilde{Y}_t)^2|X_1,\ldots,X_n\big) = O_p(\frac{1}{n}), \quad d_2(\tilde{Y}_t^+,\tilde{Y}_t)\to 0 \quad n\to\infty, \tag{15.5}$$

and

$$\mathrm{E}\big((\tilde{Y}_t - Y_t)^2|X_1,\ldots,X_n\big) \leq K\rho^t, \quad d_2(\tilde{Y}_t,Y_t)\to 0 \quad n\to\infty. \tag{15.6}$$

PROOF. We first prove (15.4). It is clear from the definitions that

$$\mathrm{E}\big((Y_t^+ - \tilde{Y}_t^+)^2|\ X_1,\ldots,X_n\big) \leq \sum_{j=0}^{t}([A(\underline{\phi})^j]_{1,1} - [A(\hat{\underline{\phi}}_n)^j]_{1,1})^2\mathrm{E}((\varepsilon_j^+)^2|X_1,\ldots,X_n). \tag{15.7}$$

Using Lemma 15.2.1 we have that $\mathrm{E}((\varepsilon_j^+)^2|X_1,\ldots,X_n)$ is the same for all j and $\mathrm{E}((\varepsilon_j^+)^2|X_1,\ldots,X_n) \xrightarrow{\mathcal{P}} \mathrm{E}(\varepsilon_t^2)$, hence we will consider for now $([A(\underline{\phi})^j]_{1,1} - [A(\hat{\underline{\phi}}_n)^j]_{1,1})^2$. Using (14.14) we have $(\hat{\underline{\phi}}_n - \underline{\phi}) = O_p(n^{-1/2})$, therefore by the mean value theorem we have $[A(\phi) - A(\hat{\underline{\phi}}_n)| = (\hat{\underline{\phi}}_n - \underline{\phi})D \approx \frac{K}{n}D$ (for some random matrix D). Hence

$$A(\hat{\underline{\phi}}_n)^j = (A(\underline{\phi}) + \frac{K}{n}D)^j = A(\underline{\phi})^j\left(1 + A(\underline{\phi})^{-1}\frac{K}{n}\right)^j$$

(note these are heuristic bounds, and this argument needs to be made precise). Applying the mean value theorem again we have

$$A(\underline{\phi})^j\left(1 + A(\underline{\phi})^{-1}\frac{K}{n}D\right)^j = A(\underline{\phi})^j + \frac{K}{n}D\,A(\underline{\phi})^j(1 + A(\underline{\phi})^{-1}\frac{K}{n}B)^j,$$

where B is such that $\|B\|_{spec} \leq \|\frac{K}{n}D\|$. Altogether this gives

$$|[A(\underline{\phi})^j - A(\hat{\underline{\phi}}_n)^j]_{1,1}| \leq \frac{K}{n} D\, A(\underline{\phi})^j (1 + A(\underline{\phi})^{-1}\frac{K}{n}B)^j.$$

Notice that for large enough n, $(1 + A(\underline{\phi})^{-1}\frac{K}{n}B)^j$ is increasing slower (as $n \to \infty$) than $A(\underline{\phi})^j$ is contracting. Therefore for a large enough n we have

$$\big|[A(\underline{\phi})^j - A(\hat{\underline{\phi}}_n)^j]_{1,1}\big| \leq \frac{K}{n^{1/2}}\rho^j,$$

for any $\frac{1}{1+\delta} < \rho < 1$. Subsituting this into (15.7) gives

$$\mathrm{E}\big((Y_t^+ - \tilde{Y}_t^+)^2 | X_1, \ldots, X_n\big) \leq \frac{K}{n^{1/2}}\mathrm{E}((\varepsilon_t^+)^2)\sum_{j=0}^{t}\rho^j = O_p(\frac{1}{n}) \to 0 \quad n \to \infty.$$

hence $d_2(\tilde{Y}_t^+, Y_t^+) \to 0$ as $n \to \infty$.

We now prove (15.5). We see that

$$\mathrm{E}\big((\tilde{Y}_t^+ - \tilde{Y}_t)^2 | X_1, \ldots, X_n\big) = \sum_{j=0}^{t} a_j^2 \mathrm{E}(\hat{\varepsilon}_{J_{t-j}} - \varepsilon_{J_{t-j}})^2 = \mathrm{E}(\hat{\varepsilon}_{J_{t-j}} - \varepsilon_{J_{t-j}})^2 \sum_{j=0}^{t} a_j^2. \tag{15.8}$$

Now by substituting (15.2) into the above we have $\mathrm{E}(\tilde{Y}_t^+ - \tilde{Y}_t)^2 = O(n^{-1})$, as required. This means that $d_2(\tilde{Y}_t^+, \tilde{Y}_t) \to 0$.

Finally we prove (15.6). We see that

$$\mathrm{E}\big((\tilde{Y}_t - Y_t)^2 | X_1, \ldots, X_n\big) = \sum_{j=t+1}^{\infty} a_j^2 \mathrm{E}(\varepsilon_t^2). \tag{15.9}$$

Using (4.21) we have $\mathrm{E}(\tilde{Y}_t - Y_t)^2 \leq K\rho^t$, thus giving us (15.6). □

We can now almost prove the result. To do this we note that

$$(\hat{\gamma}_p - \hat{\Gamma}_p\underline{\phi}) = \frac{1}{n-p}\sum_{t=p+1}^{n}\varepsilon_t \mathbf{X}_{t-1}, \quad (\gamma_p^+ - \Gamma_p^+\hat{\underline{\phi}}_n) = \frac{1}{n-p}\sum_{t=p+1}^{n}\varepsilon_t^+ \mathbf{X}_{t-1}^+. \tag{15.10}$$

Lemma 15.2.4 *Let Y_t, Y_t^+, $\tilde{Y}_t^+$ and $\tilde{Y}_t$, be defined as in Lemma 15.2.3. Define $\bar{\Gamma}_p$ and $\bar{\Gamma}_p^+$, $\bar{\gamma}_p$ and γ_p^+ in the same way as $\hat{\Gamma}_p$ and $\hat{\gamma}_p$ defined in (15.1), but using Y_t and Y_t^+ defined in Lemma*

15.2.3, rspectively, rather than X_t. We have that

$$d_2(Y_t, Y_t^+) \leq \{\mathrm{E}(Y_t - Y_t^+)^2\}^{1/2} = O_p(K(n^{-1/2} + \rho^t), \tag{15.11}$$

$$d_2(Y_t, X_t) \to 0, \quad n \to \infty, \tag{15.12}$$

and

$$d_2\big(\sqrt{n}(\bar{\gamma}_p - \bar{\Gamma}_p\underline{\phi}), \sqrt{n}(\bar{\gamma}_p^+ - \bar{\Gamma}_p^+\hat{\underline{\phi}}_n)\big) \leq n\mathrm{E}\big((\bar{\gamma}_p - \bar{\Gamma}_p\underline{\phi}) - (\bar{\gamma}_p^+ - \bar{\Gamma}_p^+\hat{\underline{\phi}}_n)\big)^2 \to 0 \quad n \to \infty, \tag{15.13}$$

where $\bar{\Gamma}_p$, $\bar{\Gamma}_p^+$, $\bar{\gamma}_p$ and $\bar{\gamma}_p^+$ are defined in the same was as $\hat{\Gamma}_p$, Γ_p^+, $\hat{\gamma}_p$ and γ_p^+, but with $\{Y_t\}$ replacing X_t in $\bar{\Gamma}_p$ and $\bar{\gamma}_p$ and $\{Y_t^+\}$ replacing X_t^+ in $\bar{\Gamma}_p^+$ and $\bar{\gamma}_p^+$. Furthermore we have

$$\mathrm{E}|\bar{\Gamma}_p^+ - \bar{\Gamma}_p| \to 0, \tag{15.14}$$

$$d_2\big((\bar{\gamma}_p - \bar{\Gamma}_p\underline{\phi}), (\gamma_p - \Gamma_p\underline{\phi})\big) \to 0, \quad \mathrm{E}|\bar{\Gamma}_p - \hat{\Gamma}_p| \to 0 \quad n \to \infty. \tag{15.15}$$

PROOF. We first prove (15.11). Using the triangle inequality we have

$$\begin{aligned}\{\mathrm{E}\big((\tilde{Y}_t - Y_t^+)^2|X_1, \ldots, X_n\big)\}^{1/2} \quad &\leq \quad \{\big(\mathrm{E}(Y_t - \tilde{Y}_t)^2|X_1, \ldots, X_n\big)\}^{1/2} + \{\big(\mathrm{E}(\tilde{Y}_t - \tilde{Y}_t^+)^2|X_1, \ldots, X_n\big)\}^{1/2} \\ &\quad + \{\mathrm{E}\big((\tilde{Y}_t^+ - Y_t^+)^2|X_1, \ldots, X_n\big)\}^{1/2} = O(n^{-1/2} + \rho^t),\end{aligned}$$

where we use Lemma 15.2.3 we get the second inequality above. Therefore by definition of $d_2(X_t, X_t^+)$ we have (15.11). To prove (15.12) we note that the only difference between Y_t and X_t is that the $\{\varepsilon_{J_k}\}$ in Y_t, is sampled from $\{\varepsilon_{p+1}, \ldots, \varepsilon_n\}$ hence sampled from F_n, where as the $\{\varepsilon_t\}_{t=p+1}^n$ in X_t are iid random variables with distribution F. Since $d_2(F_n, F) \to 0$ (**?**, Lemma 8.4) it follows that $d_2(Y_t, X_t) \to 0$, thus proving (15.12).

To prove (15.13) we consider the difference $(\bar{\gamma}_p - \bar{\Gamma}_p\underline{\phi}) - (\bar{\gamma}_p^+ - \bar{\Gamma}_p^+\hat{\underline{\phi}}_n)$ and use (15.10) to get

$$\frac{1}{n}\sum_{t=p+1}^n \Big\{\varepsilon_t\mathbf{Y}_{t-1} - \varepsilon_t^+\mathbf{Y}_{t-1}^+\Big\} = \frac{1}{n}\sum_{t=p+1}^n \Big\{(\varepsilon_t - \varepsilon_t^+)\mathbf{Y}_{t-1} + \varepsilon_t^+(\mathbf{Y}_{t-1} - \mathbf{Y}_{t-1}^+)\Big\},$$

where we note that $\mathbf{Y}_{t-1}^+ = (Y_{t-1}^+, \ldots, Y_{t-p}^+)'$ and $\mathbf{Y}_{t-1} = (Y_{t-1}, \ldots, Y_{t-p})'$. Using the above,

and taking conditional expectations with respect to $\{X_1, \ldots, X_n\}$ and noting that conditioned on $\{X_1, \ldots, X_n\}$, $(\varepsilon_t - \varepsilon_t^+)$ are independent of $\mathbf{X}_k$ and $\mathbf{X}_k^+$ for $k < t$ we have

$$\left\{\frac{1}{n}\mathrm{E}\Big(\sum_{t=p+1}^{n}\left\{\varepsilon_t\mathbf{Y}_{t-1} - \varepsilon_t^+\mathbf{Y}_{t-1}^+\right\}\Big)^2 |X_1, \ldots, X_n\right\}^{1/2} \leq I + II$$

where

$$\begin{aligned}
I &= \frac{1}{n}\sum_{t=p+1}^{n}\{\mathrm{E}\big((\varepsilon_t - \varepsilon_t^+)^2|X_1, \ldots, X_n\big)\}^{1/2}\{\mathrm{E}(\mathbf{Y}_{t-1}^2|X_1, \ldots, X_n)\}^{1/2} \\
&= \{\mathrm{E}\big((\varepsilon_t - \varepsilon_t^+)^2|X_1, \ldots, X_n\big)\}^{1/2}\frac{1}{n}\sum_{t=p+1}^{n}\{\mathrm{E}((\mathbf{Y}_{t-1}^2|X_1, \ldots, X_n)\}^{1/2} \\
II &= \frac{1}{n}\sum_{t=p+1}^{n}\{\mathrm{E}((\varepsilon_t^+)^2|X_1, \ldots, X_n)\}^{1/2}\{\mathrm{E}((\mathbf{Y}_{t-1} - \mathbf{Y}_{t-1}^+)^2|X_1, \ldots, X_n\}^{1/2} \\
&= \{\mathrm{E}((\varepsilon_t^+)^2|X_1, \ldots, X_n)\}^{1/2}\frac{1}{n}\sum_{t=p+1}^{n}\{\mathrm{E}((\mathbf{Y}_{t-1} - \mathbf{Y}_{t-1}^+)^2|X_1, \ldots, X_n)\}^{1/2}.
\end{aligned}$$

Now by using (15.2) we have $I \leq Kn^{-1/2}$, and (15.13) and Corollary 15.2.1 we obtain $II \leq Kn^{-1/2}$, hence we have (15.13). Using a similar technique to that given above we can prove (15.14).

(15.15) follows from (15.13), (15.14) and (15.12). □

Corollary 15.2.2 *Let Γ_p^+, $\hat{\Gamma}_p$, $\hat{\gamma}_p$ and γ_p^+ be defined in (15.1). Then we have*

$$d_2\left(\sqrt{n}(\hat{\gamma}_p - \hat{\Gamma}_p\underline{\phi}), \sqrt{n}(\gamma_p^+ - \Gamma_p^+\hat{\underline{\phi}}_n)\right) \to 0 \tag{15.16}$$

$$d_1(\Gamma_p^+, \hat{\Gamma}_p) \to 0, \tag{15.17}$$

as $n \to \infty$.

PROOF. We first prove (15.16). Using (15.13), (15.15) and the triangular inequality gives (15.16). To prove (15.17) we use (15.14) and (15.15) and the triangular inequality and (15.16) immediately follows. □

Now by using (15.17) and Lemma 15.2.1 we have

$$\Gamma_p^+ \xrightarrow{\mathcal{P}} \mathrm{E}(\Gamma_p),$$

and by using (15.16), the distribution of $\sqrt{n}(\gamma_p^+ - \Gamma_p^+ \hat{\underline{\phi}}_n$ converges weakly to the distribution of $\sqrt{n}(\hat{\gamma}_p - \hat{\Gamma}_p \underline{\phi})$. Therefore

$$\sqrt{n}(\hat{\underline{\phi}}_n^+ - \hat{\underline{\phi}}_n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\Gamma_p^{-1}),$$

hence the distributions of $\sqrt{n}(\hat{\gamma}_p - \hat{\Gamma}_p \underline{\phi})$ and $\sqrt{n}(\gamma_p^+ - \Gamma_p^+ \hat{\underline{\phi}}_n)$ aymptotically coincide. From (14.15) we have $\sqrt{n}(\hat{\underline{\phi}}_n - \underline{\phi}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2 \Gamma_p^{-1})$. Thus we see that the distribution of $\sqrt{n}(\hat{\underline{\phi}}_n - \underline{\phi})$ and $\sqrt{n}(\hat{\underline{\phi}}_n^+ - \hat{\underline{\phi}}_n)$ asymptotically coincide.

Appendix A

Background

A.1 Some definitions and inequalities

- Some norm definitions.

  The norm of an object, is a postive numbers which measure the 'magnitude' of that object. Suppose $\underline{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, then we define $\|\underline{x}\|_1 = \sum_{j=1}^n |x_j|$ and $\|\underline{x}\|_2 = (\sum_{j=1}^n |x_j^2)^{1/2}$ (this is known as the Euclidean norm). There are various norms for matrices, the most popular is the spectral norm $\|\cdot\|_{spec}$: let A be a matrix, then $\|A\|_{spec} = \lambda_{max}(AA')$, where λ_{max} denotes the largest eigenvalue.

- $\mathbb{Z}$ denotes the set of a integers $\{\ldots, -1, 0, 1, 2, \ldots\}$. $\mathbb{R}$ denotes the real line $(-\infty, \infty)$.

- Complex variables.

  $i = \sqrt{-1}$ and the complex variable $z = x + iy$, where x and y are real.

  Often the radians representation of a complex variable is useful. If $z = x + iy$, then it can also be written as $r\exp(i\theta)$, where $r = \sqrt{x^2 + y^2}$ and $\theta = \tan^{-1}(y/x)$.

  If $z = x + iy$, its complex conjugate is $\bar{z} = x - iy$.

- The roots of a rth order polynomial $a(z)$, are those values $\lambda_1, \ldots, \lambda_r$ where $a(\lambda_i) = 0$ for $i = 1, \ldots, r$.

- Let $\lambda(A)$ denote the spectral radius of the the matrix A (the largest eigenvalue in absolute terms). Then for any matrix norm $\|A\|$ we have $\lim_{j\to\infty} \|A^j\|^{1/j} = \lambda(A)$ (see Gelfand's

formula). Suppose $\lambda(A) < 1$, then Gelfand's formula implies that for any $\lambda(A) < \rho < 1$, there exists a constant, C, (which only depends A and ρ), such that $\|A^j\| \leq C_{A,\rho}\rho^j$.

- The mean value theorem.

  This basically states that if the partial derivative of the function $f(x_1, x_2, \ldots, x_n)$ has a bounded in the domiain Ω, then for $\underline{x} = (x_1, \ldots, x_n)$ and $\underline{y} = (y_1, \ldots, y_n)$

$$f(x_1, x_2, \ldots, x_n) - f(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} (x_i - y_i) \frac{\partial f}{\partial x_i} \rfloor_{\underline{x}=\underline{x}^*}$$

  where $\underline{x}^*$ lies somewhere between $\underline{x}$ and $\underline{y}$.

- The Taylor series expansion.

  This is closely related to the mean value theorem and a second order expansion is

$$f(x_1, x_2, \ldots, x_n) - f(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} (x_i - y_i) \frac{\partial f}{\partial x_i} + \sum_{i,j=1}^{n} (x_i - y_i)(x_j - y_j) \frac{\partial f^2}{\partial x_i \partial x_j} \rfloor_{\underline{x}=\underline{x}^*}$$

- Partial Fractions.

  We use the following result mainly for obtaining the MA(∞) expansion of an AR process.

  Suppose that $|g_i| > 1$ for $1 \leq i \leq n$. Then if $g(z) = \prod_{i=1}^{n} (1 - z/g_i)^{r_i}$, the inverse of $g(z)$ satisfies

$$\frac{1}{g(z)} = \sum_{i=1}^{n} \{ \sum_{j=1}^{r_i} \frac{g_{i,j}}{(1 - \frac{z}{g_i})^j} \},$$

  where $g_{i,j} =$ ..... Now we can make a polynomial series expansion of $(1 - \frac{z}{g_i})^{-j}$ which is valid for all $|z| \leq 1$.

- Dominated convergence.

  Suppose a sequence of functions $f_n(x)$ is such that pointwise $f_n(x) \to f(x)$ and for all n and x, $|f_n(x)| \leq g(x)$, then $\int f_n(x)dx \to \int f(x)dx$ as $n \to \infty$.

  We use this result all over the place to exchange infinite sums and expectations. For example,

if $\sum_{j=1}^{\infty}|a_j|\mathrm{E}(|Z_j|)<\infty$, then by using dominated convergence we have

$$\mathrm{E}(\sum_{j=1}^{\infty}a_jZ_j)=\sum_{j=1}^{\infty}a_j\mathrm{E}(Z_j).$$

- Dominated convergence can be used to prove the following lemma. A more hands on proof is given below the lemma.

Lemma A.1.1 *Suppose* $\sum_{k=-\infty}^{\infty}|c(k)|<\infty$, *then we have*

$$\frac{1}{n}\sum_{k=-(n-1)}^{(n-1)}|kc(k)|\rightarrow 0$$

as $n\rightarrow\infty$. *Moreover, if* $\sum_{k=-\infty}^{\infty}|kc(k)|<\infty$, *then* $\frac{1}{n}\sum_{k=-(n-1)}^{(n-1)}|kc(k)|=O(\frac{1}{n})$.

PROOF. The proof is straightforward in the case that $\sum_{k=\infty}^{\infty}|kc(k)|<\infty$ (the second assertion), in this case $\sum_{k=-(n-1)}^{(n-1)}\frac{|k|}{n}|c(k)|=O(\frac{1}{n})$. The proof is slightly more tricky in the case that $\sum_{k=\infty}^{\infty}|c(k)|<\infty$. First we note that since $\sum_{k=-\infty}^{\infty}|c(k)|<\infty$ for every $\varepsilon>0$ there exists a N_ε such that for all $n\geq N_\varepsilon$, $\sum_{|k|\geq n}|c(k)|<\varepsilon$. Let us suppose that $n>N_\varepsilon$, then we have the bound

$$\begin{aligned}\frac{1}{n}\sum_{k=-(n-1)}^{(n-1)}|kc(k)| &\leq \frac{1}{n}\sum_{k=-(N_\varepsilon-1)}^{(N_\varepsilon-1)}|kc(k)|+\frac{1}{n}\sum_{N_\varepsilon\leq|k|\leq n}|kc(k)|\\ &\leq \frac{1}{2\pi n}\sum_{k=-(N_\varepsilon-1)}^{(N_\varepsilon-1)}|kc(k)|+\varepsilon.\end{aligned}$$

Hence if we keep N_ε fixed we see that $\frac{1}{n}\sum_{k=-(N_\varepsilon-1)}^{(N_\varepsilon-1)}|kc(k)|\rightarrow 0$ as $n\rightarrow\infty$. Since this is true for all ε (for different thresholds N_ε) we obtain the required result. □

- Cauchy Schwarz inequality.

  In terms of sequences it is

$$|\sum_{j=1}^{\infty}a_jb_j|\leq(\sum_{j=1}^{\infty}a_j^2)^{1/2}(\sum_{j=1}^{\infty}b_j^2)^{1/2}$$

. For integrals and expectations it is

$$\mathrm{E}|XY| \leq \mathrm{E}(X^2)^{1/2}\mathrm{E}(Y^2)^{1/2}$$

- Holder's inequality.

  This is a generalisation of the Cauchy Schwarz inequality. It states that if $1 \leq p, q \leq \infty$ and $p + q = 1$, then

  $$\mathrm{E}|XY| \leq \mathrm{E}(|X|^p)^{1/p}\mathrm{E}(|Y|^q)^{1/q}$$

  . A similar results is true for sequences too.

- Martingale differences. Let $\mathcal{F}_t$ be a sigma-algebra, where $X_t, X_{t-1}, \ldots \in \mathcal{F}_t$. Then $\{X_t\}$ is a sequence of martingale differences if $\mathrm{E}(X_t|\mathcal{F}_{t-1}) = 0$.

- Minkowski's inequality.

  If $1 < p < \infty$, then

  $$(\mathrm{E}(\sum_{i=1}^{n} X_i)^p)^{1/p} \leq \sum_{i=1}^{n}(\mathrm{E}(|X_i|^p))^{1/p}.$$

- Doob's inequality.

  This inequality concerns martingale differences. Let $\mathcal{S}_n = \sum_{t=1}^{n} X_t$, then

  $$\mathrm{E}(\sup_{n \leq N} |\mathcal{S}_n|^2) \leq \mathrm{E}(\mathcal{S}_N^2).$$

- Burkhölder's inequality.

  Suppose that $\{X_t\}$ are martingale differences and define $S_n = \sum_{k=1}^{n} X_t$. For any $p \geq 2$ we have

  $$\{\mathrm{E}(S_n^p)\}^{1/p} \leq \left(2p \sum_{k=1}^{n} \mathrm{E}(X_k^p)^{2/p}\right)^{1/2}.$$

  An application, is to the case that $\{X_t\}$ are identically distributed random variables, then we have the bound $\mathrm{E}(S_n^p) \leq \mathrm{E}(X_0^p)^2(2p)^{p/2}n^{p/2}$.

  It is worthing noting that the Burkhölder inequality can also be defined for $p < 2$ (see

Davidson (1994), pages 242). It can also be generalised to random variables $\{X_t\}$ which are not necessarily martingale differences (see Dedecker and Doukhan (2003)).

- Riemann-Stieltjes Integrals.

  In basic calculus we often use the basic definition of the Riemann integral, $\int g(x)f(x)dx$, and if the function $F(x)$ is continuous and $F'(x) = f(x)$, we can write $\int g(x)f(x)dx = \int g(x)dF(x)$. There are several instances where we need to broaden this definition to include functions F which are not continuous everywhere. To do this we define the Riemann-Stieltjes integral, which coincides with the Riemann integral in the case that $F(x)$ is continuous.

  $\int g(x)dF(x)$ is defined in a slightly different way to the Riemann integral $\int g(x)f(x)dx$. Let us first consider the case that $F(x)$ is the step function $F(x) = \sum_{i=1}^{n} a_i I_{[x_{i-1},x_i]}$, then $\int g(x)dF(x)$ is defined as $\int g(x)dF(x) = \sum_{i=1}^{n}(a_i - a_{i-1})g(x_i)$ (with $a_{-1} = 0$). Already we see the advantage of this definition, since the derivative of the step function is not well defined at the jumps. As most functions can be written as the limit of step functions ($F(x) = \lim_{k\infty} F_k(x)$, where $F_k(x) = \sum_{i=1}^{n_k} a_{i,n_k} I_{[x_{i_{k-1}-1},x_{i_k}]}$), we define $\int g(x)dF(x) = \lim_{k\to\infty} \sum_{i=1}^{n_k}(a_{i,n_k} - a_{i-1,n_k})g(x_{i_k})$.

  In statistics, the function F will usually be non-decreasing and bounded. We call such functions distributions.

Theorem A.1.1 (Helly's Theorem) *Suppose that $\{F_n\}$ are a sequence of distributions with $F_n(-\infty) = 0$ and $\sup_n F_n(\infty) \leq M < \infty$. There exists a distribution F, and a subsequence F_{n_k} such that for each $x \in \mathbb{R}$ $F_{n_k} \to F$ and F is right continuous.*

A.2 Martingales

Definition A.2.1 *A sequence $\{X_t\}$ is said to be a martingale difference if $\mathrm{E}[X_t|\mathcal{F}_{t-1}]$, where $\mathcal{F}_{t=1} = \sigma(X_{t-1}, X_{t-2}, \ldots)$. In other words, the best predictor of X_t given the past is simply zero.*

Martingales are very useful when proving several results, including central limit theorems.

Martingales arise naturally in several situations. We now show that if correct likelihood is used (not the quasi-case), then the gradient of the conditional log likelihood evaluated at the true parameter is the sum of martingale differences. To see why, let $\mathcal{B}_T = \sum_{t=2}^{T} \log f_\theta(X_t|X_{t-1}, \ldots, X_1)$

be the conditonal log likelihood and $\mathcal{C}_T(\theta)$ its derivative, where

$$\mathcal{C}_T(\theta) = \sum_{t=2}^{T} \frac{\partial \log f_\theta(X_t|X_{t-1},\ldots,X_1)}{\partial\theta}.$$

We want to show that $\mathcal{C}_T(\theta_0)$ is the sum of martingale differences. By definition if $\mathcal{C}_T(\theta_0)$ is the sum of martingale differences then

$$\mathrm{E}\left(\frac{\partial \log f_\theta(X_t|X_{t-1},\ldots,X_1)}{\partial\theta}\rfloor_{\theta=\theta_0}\middle| X_{t-1}, X_{t-2},\ldots,X_1\right) = 0,$$

we will show this. Rewriting the above in terms of integrals and exchanging derivative with integral we have

$$\begin{aligned}
& \mathrm{E}\left(\frac{\partial \log f_\theta(X_t|X_{t-1},\ldots,X_1)}{\partial\theta}\rfloor_{\theta=\theta_0}\middle| X_{t-1}, X_{t-2},\ldots,X_1\right) \\
= & \int \frac{\partial \log f_\theta(x_t|X_{t-1},\ldots,X_1)}{\partial\theta}\rfloor_{\theta=\theta_0} f_{\theta_0}(x_t|X_{t-1},\ldots,X_1)dx_t \\
= & \int \frac{1}{f_{\theta_0}(x_t|X_{t-1},\ldots,X_1)}\frac{\partial f_\theta(x_t|X_{t-1},\ldots,X_1)}{\partial\theta}\rfloor_{\theta=\theta_0} f_{\theta_0}(x_t|X_{t-1},\ldots,X_1)dx_t \\
= & \frac{\partial}{\partial\theta}\left(\int f_\theta(x_t|X_{t-1},\ldots,X_1)dx_t\right)\rfloor_{\theta=\theta_0} = 0.
\end{aligned}$$

Therefore $\{\frac{\partial \log f_\theta(X_t|X_{t-1},\ldots,X_1)}{\partial\theta}\rfloor_{\theta=\theta_0}\}_t$ are a sequence of martingale differences and $\mathcal{C}_t(\theta_0)$ is the sum of martingale differences (hence it is a martingale).

A.3 The Fourier series

The Fourier transform is a commonly used tool. We recall that $\{\exp(2\pi ij\omega); j \in \mathbb{Z}\}$ is an orthogonal basis of the space $L^2[0,1]$. In other words, if $f \in L^2[0,1]$ (ie, $\int_0^2 f(\omega)^2 d\omega < \infty$) then

$$f_n(u) = \sum_{j=-n}^{n} c_j e^{iju2\pi} \qquad c_j = \int_0^1 f(u)\exp(i2\pi ju)du,$$

where $\int |f(u) - f_n(u)|^2 du \to 0$ as $n \to \infty$. Roughly speaking, if the function is continuous then we can say that

$$f(u) = \sum_{j\in\mathbb{Z}} c_j e^{iju}.$$

An important property is that $f(u) \equiv$constant iff $c_j = 0$ for all $j \neq 0$. Moreover, for all $n \in \mathbb{Z}$ $f(u+n) = f(u)$ (hence f is periodic).

Some relations:

(i) **Discrete Fourier transforms of finite sequences**

It is straightforward to show (by using the property $\sum_{j=1}^{n} \exp(i2\pi k/n) = 0$ for $k \neq 0$) that if

$$d_k = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} x_j \exp(i2\pi jk/n),$$

then $\{x_r\}$ can be recovered by inverting this transformation

$$x_r = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} d_k \exp(-i2\pi rk/n),$$

(ii) **Fourier sums and integrals**

Of course the above only has meaning when $\{x_k\}$ is a finite sequence. However suppose that $\{x_k\}$ is a sequence which belongs to ℓ_2 (that is $\sum_k x_k^2 < \infty$), then we can define the function

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

where $\int_0^{2\pi} f(\omega)^2 d\omega = \sum_k x_k^2$, and we we can recover $\{x_k\}$ from $f(\omega)$, through

$$x_k = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} f(\omega) \exp(-ik\omega).$$

(iii) **Convolutions**. Let us suppose that $\sum_k |a_k|^2 < \infty$ and $\sum_k |b_k|^2 < \infty$ and we define the Fourier transform of the sequences $\{a_k\}$ and $\{b_k\}$ as $A(\omega) = \frac{1}{\sqrt{2\pi}} a_k \exp(ik\omega)$ and $B(\omega) = \frac{1}{\sqrt{2\pi}} \sum_k b_k \exp(ik\omega)$ respectively. Then

$$\begin{aligned} \sum_{j=-\infty}^{\infty} a_j b_{k-j} &= \int_0^{2\pi} A(\omega) B(-\omega) \exp(-ik\omega) d\omega \\ \sum_{j=-\infty}^{\infty} a_j b_j \exp(ij\omega) &= \int_0^{2\pi} A(\lambda) B(\omega - \lambda) d\lambda. \end{aligned} \tag{A.1}$$

The proof of the above follows from

$$\begin{aligned}\sum_{j=-\infty}^{\infty} a_j b_j \exp(ij\omega) &= \sum_{r=-\infty}^{\infty} \int_0^{2\pi}\int_0^{2\pi} A(\lambda_1)B(\lambda_2)\exp(-ir(\lambda_1+\lambda_2))\exp(ij\omega) \\ &= \int\int A(\lambda_1)B(\lambda_2) \underbrace{\sum_{r=-\infty}^{\infty} \exp(ir(\omega-\lambda_1-\lambda_2))}_{=\delta_\omega(\lambda_1+\lambda_2)} d\lambda_1 d\lambda_2 \\ &= \int_0^{2\pi} A(\lambda)B(\omega-\lambda)d\lambda.\end{aligned}$$

(iv) **Using the DFT to calculate convolutions**. Our objective is calculate $\sum_{j=k}^{n} a_j b_{j-s}$ for all $s = 0, \ldots, n-1$ in as few computing computing operations. This is typically done via the DFT. Examples in time series where this is useful is in calculating the sample autocovariance function.

Suppose we have two sequences $\underline{a} = (a_1, \ldots, a_n)$ and $\underline{b} = (b_1, \ldots, b_n)$. Let $A_n(\omega_{k,n}) = \sum_{j=1}^{n} a_j \exp(ij\omega_{k,n})$ and $B_n(\omega_{k,n}) = \sum_{j=1}^{n} b_j \exp(ij\omega_{k,n})$ where $\omega_{k,n} = 2\pi k/n$. It is straightforward to show that

$$\frac{1}{n}\sum_{k=1}^{n} A_n(\omega_{k,n})\overline{B_n(\omega_{k,n})}\exp(-is\omega_{k,n}) = \sum_{j=s}^{n} a_j b_{j-s} + \sum_{j=1}^{s-1} a_j b_{j-s+n},$$

this is very fast to compute (requiring only $O(n\log n)$ operations using first the FFT and then inverse FFT). The only problem is that we don't want the second term.

By padding the sequences and defining $A_n(\omega_{k,2n}) = \sum_{j=1}^{n} a_j \exp(ij\omega_{k,2n}) = \sum_{j=1}^{2n} a_j \exp(ij\omega_{k,2n})$, with $\omega_{k,2n} = 2\pi k/2n$ (where we set $a_j = 0$ for $j > 0$) and analogously $B_n(\omega_{k,2n}) = \sum_{j=1}^{n} b_j \exp(ij\omega_{k,2n})$, we are able to remove the second term. Using the same calculations we have

$$\frac{1}{n}\sum_{k=1}^{2n} A_n(\omega_{k,2n})\overline{B_n(\omega_{k,2n})}\exp(-is\omega_{k,2n}) = \sum_{j=s}^{n} a_j b_{j-s} + \underbrace{\sum_{j=1}^{s-1} a_j b_{j-s+2n}}_{=0}.$$

This only requires $O(2n\log(2n))$ operations to compute the convolution for all $0 \le k \le n-1$.

(v) **The Poisson Summation Formula** Suppose we do not observe the entire function and observe a sample from it, say $f_{t,n} = f(\frac{t}{n})$ we can use this to estimate the Fourier coefficient

c_j via the Discrete Fourier Transform:

$$c_{j,n} = \frac{1}{n}\sum_{t=1}^{n} f(\frac{t}{n})\exp(ij\frac{2\pi t}{n}).$$

The *Poisson Summation formula* is

$$c_{j,n} = c_j + \sum_{k=1}^{\infty} c_{j+kn} + \sum_{k=1}^{\infty} c_{j-kn},$$

which we can prove by replacing $f(\frac{t}{n})$ with $\sum_{j\in\mathbb{Z}} c_j e^{ij2\pi t/n}$. In other words, $c_{j,n}$ cannot disentangle frequency $e^{ij\omega}$ from it's harmonics $e^{i(j+n)\omega}$ (this is *aliasing*).

(vi) **Error in the DFT** By using the Poisson summation formula we can see that

$$|c_{j,n} - c_j| \leq \sum_{k=1}^{\infty} |c_{j+kn}| + \sum_{k=1}^{\infty} |c_{j-kn}|$$

It can be shown that if a function $f(\cdot)$ is $(p+1)$ times differentiable with bounded derivatives or that $f^p(\cdot)$ is bounded and piecewise montonic then the corresponding Fourier coefficients satisfy

$$|c_j| \leq C|j|^{-(p+1)}.$$

Using this result and the Poisson summation formula we can show that for $|j| \leq n/2$ that if if a function $f(\cdot)$ is $(p+1)$ times differentiable with bounded derivatives or that $f^p(\cdot)$ is piecewise montonic and $p \geq 1$ then

$$|c_{j,n} - c_j| \leq Cn^{-(p+1)}, \tag{A.2}$$

where C is some finite constant. However, we cannot use this result in the case that f is bounded and piecewise monotone, however it can still be shown that

$$|c_{j,n} - c_j| \leq Cn^{-1}, \tag{A.3}$$

see Section 6.3, page 189, Briggs and Henson (1997).

A.4 Application of Burkholder's inequality

There are two inequalities (one for $1 < p \leq 2$). Which is the following:

Theorem A.4.1 *Suppose that Y_k are martingale differences and that $S_n = \sum_{j=1}^{n} Y_k$, then for $0 < q \leq 2$*

$$\mathrm{E}|S_n|^q \leq 2 \sum_{j=1}^{n} \mathrm{E}(X_k^q), \tag{A.4}$$

See for example Davidson (p. 242, Theorem 15.17).

And one for ($p \geq 2$), this is the statement for the Burkölder inequality:

Theorem A.4.2 *Suppose $\{S_i : \mathcal{F}_i\}$ is a martingale and $1 < p < \infty$. Then there exists constants C_1, C_2 depending only on p such that*

$$C_1 \mathrm{E}\left(\sum_{i=1}^{m} X_i^2\right)^{p/2} \leq \mathrm{E}|S_n|^p \leq C_2 \mathrm{E}\left(\sum_{i=1}^{m} X_i^2\right)^{p/2}. \tag{A.5}$$

An immediately consequence of the above for $p \geq 2$ is the following corollary (by using Hölder's inequality):

Corollary A.4.1 *Suppose $\{S_i : \mathcal{F}_i\}$ is a martingale and $2 \leq p < \infty$. Then there exists constants C_1, C_2 depending only on p such that*

$$\|S_n\|_p^E \leq \left(C_2^{2/p} \sum_{i=1}^{m} \|X_i^2\|_{p/2}^E\right)^{1/2}. \tag{A.6}$$

PROOF. By using the right hand side of (A.5) we have

$$\begin{aligned} \left\{\mathrm{E}|S_n|^p\right\}^{1/p} &\leq \left[\left(C_2 \mathrm{E}\left(\sum_{i=1}^{m} X_i^2\right)^{p/2}\right)^{2/p}\right]^{1/2} \\ &= \left[C_2^{2/p} \left\|\sum_{i=1}^{m} X_i^2\right\|_{p/2}^E\right]^{1/2}. \end{aligned} \tag{A.7}$$

By using Hölder inequality we have

$$\left\{\mathrm{E}|S_n|^p\right\}^{1/p} \leq \left[C_2^{2/p}\sum_{i=1}^{m}\|X_i^2\|_{p/2}^E\right]^{1/2}. \tag{A.8}$$

Thus we have the desired result. □

We see the value of the above result in the following application. Suppose $S_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ and $\|X_k\|_p^E \leq K$. Then we have

$$\begin{aligned}\mathrm{E}\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)^p &\leq \left[\frac{1}{n}C_2^{2/p}\sum_{k=1}^{n}\|X_k^2\|_{p/2}^E\right]^{p/2} \\ &\leq \frac{C_2}{n^p}\left[\sum_{k=1}^{n}\|X_k^2\|_{p/2}^E\right]^{p/2} \leq \frac{C_2}{n^p}\left[\sum_{k=1}^{n}K^2\right]^{p/2} = O(\frac{1}{n^{p/2}}).\end{aligned} \tag{A.9}$$

Below is the result that that Moulines et al (2004) use (they call it the generalised Burkholder inequality) the proof can be found in Dedecker and Doukhan (2003). Note that it is for $p \geq 2$, which I forgot to state in what I gave you.

Lemma A.4.1 *Suppose* $\{\phi_k : k = 1, 2, \ldots\}$ *is a stochastic process which satisfies* $\mathrm{E}(\phi_k) = 0$ *and* $\mathrm{E}(\phi_k^p) < \infty$ *for some* $p \geq 2$. *Let* $\mathcal{F}_k = \sigma(\phi_k, \phi_{k-1}, \ldots)$. *Then we have that*

$$\left\|\sum_{k=1}^{s}\phi_k\right\|_p^E \leq \left(2p\sum_{k=1}^{s}\|\phi_k\|_p^E\sum_{j=k}^{s}\|\mathrm{E}(\phi_j|\mathcal{F}_k)\|_p^E\right)^{1/2}. \tag{A.10}$$

We note if $\sum_{j=k}^{s}\|\mathrm{E}(\phi_j|\mathcal{F}_k)\|_p^E < \infty$, then we (A.11) is very similar to (A.6), and gives the same rate as (A.9).

But I think one can obtain something similar for $1 \leq p \leq 2$. I think the below is correct.

Lemma A.4.2 *Suppose* $\{\phi_k : k = 1, 2, \ldots\}$ *is a stochastic process which satisfies* $\mathrm{E}(\phi_k) = 0$ *and* $\mathrm{E}(\phi_k^q) < \infty$ *for some* $1 < q \leq 2$. *Let* $\mathcal{F}_k = \sigma(\phi_k, \phi_{k-1}, \ldots)$. *Further, we suppose that there exists a* $0 < \rho < 1$, *and* $0 < K < \infty$ *such that* $\|\mathrm{E}(\phi_t|\mathcal{F}_{t-j})\|_q < K\rho^j$. *Then we have that*

$$\left\|\sum_{k=1}^{s}a_k\phi_k\right\|_q^E \leq \frac{K^*}{1-\rho}\left(\sum_{k=1}^{s}|a_k|^q\right)^{1/q}, \tag{A.11}$$

where K^* is a finite constant.

PROOF. Let $\mathrm{E}_j(\phi_k) = \mathrm{E}(\phi_k|\mathcal{F}_{k-j})$. We note that by definition $\{\phi_k\}$ is a mixingale (see, for example, Davidson (1997), chapter 16), therefore amost surely ϕ_k satisfies the representation

$$\phi_k = \sum_{j=0}^{\infty}[\mathrm{E}_{k-j}(\phi_k) - \mathrm{E}_{k-j-1}(\phi_k)]. \tag{A.12}$$

By substituting the above into the sum $\sum_{k=1}^{s} a_k\phi_k$ we obtain

$$\sum_{k=1}^{s} a_k\phi_k = \sum_{k=1}^{s}\sum_{j=0}^{\infty}[\mathrm{E}_{k-j}(\phi_k) - \mathrm{E}_{k-j-1}(\phi_k)] = \sum_{j=0}^{\infty}\left(\sum_{k=1}^{s}[\mathrm{E}_{k-j}(\phi_k) - \mathrm{E}_{k-j-1}(\phi_k)]\right). \tag{A.13}$$

Keeping j constant, we see that $\{\mathrm{E}_{k-j}(\phi_k) - \mathrm{E}_{k-j-1}(\phi_k)\}_k$ is a martingale sequence. Hence $\sum_{k=1}^{s}[\mathrm{E}_{k-j}(\phi_k) - \mathrm{E}_{k-j-1}(\phi_k)]$ is the sum of martingale differences. This implies we can apply (A.4) to (A.13), and get

$$\begin{aligned}\left\|\sum_{k=1}^{s} a_k\phi_k\right\|_q^E &\leq \sum_{j=0}^{\infty}\left\|\sum_{k=1}^{s}|a_k|[\mathrm{E}_{k-j}(\phi_k) - \mathrm{E}_{k-j-1}(\phi_k)]\right\|_q^E \\ &\leq \sum_{j=0}^{\infty}\left(2\sum_{k=1}^{s}|a_k|(\|\mathrm{E}_{k-j}(\phi_k) - \mathrm{E}_{k-j-1}(\phi_k)\|_q^E)^q\right)^{1/q}\end{aligned}$$

Under the stated assumption $\|\mathrm{E}_{k-j}(\phi_k) - \mathrm{E}_{k-j-1}(\phi_k)\|_q^E \leq 2K\rho^j$. Substituting this inequality into the above gives

$$\left\|\sum_{k=1}^{s} a_k\phi_k\right\|_q^E \leq \sum_{j=0}^{\infty}\left(2\sum_{k=1}^{s}|a_k|^q(2K\rho^j)^q\right)^{1/q} \leq 2^{1+1/q}K\sum_{j=0}^{\infty}\rho^j\left(\sum_{k=1}^{s}|a_k|^q\right)^{1/q}.$$

Thus we obtain the desired result. □

A.5 The Fast Fourier Transform (FFT)

The Discrete Fourier transform is used widely in several disciplines. Even in areas its use may not be immediately obvious (such as inverting Toeplitz matrices) it is still used because it can be evalated in a speedy fashion using what is commonly called the fast fourier transform (FFT). It is an algorithm which simplifies the number of computing operations required to compute the Fourier

transform of a sequence of data. Given that we are in the age of 'big data' it is useful to learn what one of most popular computing algorithms since the 60s actually does.

Recalling the notation in Section **??** the Fourier transform is the linear transformation

$$F_n \underline{X}_n = (J_n(\omega_0), \ldots, J_n(\omega_{n-1})).$$

If this was done without any using any tricks this requires $O(n^2)$ computing operations. By using some neat factorizations, the fft reduces this to $n \log n$ computing operations.

To prove this result we will ignore the standardization factor $(2\pi n)^{-1/2}$ and consider just the Fourier transform

$$d(\omega_{k,n}) = \underbrace{\sum_{t=1}^{n} x_t \exp(it\omega_{k,n})}_{k \text{ different frequencies}},$$

where $\omega_{k,n} = \frac{2\pi k}{n}$. Here we consider the proof for general n, later in Example A.5.1 we consider the specific case that $n = 2^m$. Let us assume that n is not a prime (if it is then we simply pad the vector with one zero and increase the length to $n+1$), then it can be factorized as $n = pq$. Using these factors we write t as $t = t_1 p + t \text{mod} p$ where t_1 is some integer value that lies between 0 to $q-1$ and $t_0 = t \text{mod} p$ lies between 0 to $p-1$. Substituting this into $d(\omega_k)$ gives

$$\begin{aligned} d(\omega_k) &= \sum_{t=1}^{n} x_t \exp[i(t_1 p + t \text{mod} p)\omega_{k,n}] \\ &= \sum_{t_0=0}^{p-1} \sum_{t_1=0}^{q-1} x_{t_1 p + t_0} \exp[i(t_1 p + t_0)\omega_{k,n}] = \sum_{t_0=0}^{p-1} \exp[it_0\omega_{k,n}] \sum_{t_1=0}^{q-1} x_{t_1 p + t_0} \exp[it_1 p \omega_{k,n}] \end{aligned}$$

It is straightforward to see that $t_1 p \omega_{k,n} = \frac{2\pi t_1 p k}{n} = \frac{2\pi t_1 k}{q} = t_1 \omega_{k,q}$ and that $\exp(it_1 p \omega_{k,n}) =$

$\exp(it_1\omega_{k,q}) = \exp(it_1\omega_{k\text{mod}q,q})$. This means $d(\omega_k)$ can be simplified as

$$\begin{aligned} d(\omega_k) &= \sum_{t_0=0}^{p-1} \exp\left[it_0\omega_{k,n}\right] \sum_{t_1=0}^{q-1} x_{t_1p+t_0} \exp\left[it_1\omega_{k\text{mod}q,q}\right] \\ &= \sum_{t_0=0}^{p-1} \exp\left[it_0\omega_{k,n}\right] \underbrace{\sum_{t_1=0}^{q-1} x_{t_1p+t_0} \exp\left[it_1\omega_{k_0,q}\right]}_{\text{embedded Fourier transform}} \\ &= \sum_{t_0=0}^{p-1} \exp\left[it_0\omega_{k,n}\right] \underbrace{A(t_0, k\text{mod}q)}_{q \text{ frequencies}}, \end{aligned}$$

where $k_0 = k\text{mod}q$ can take values from $0, \ldots, q-1$. Thus to evaluate $d(\omega_k)$ we need to evaluate $A(t_0, k\text{mod}q)$ for $0 \leq t_0 \leq p-1$, $0 \leq k_0 \leq q-1$. To evaluate $A(t_0, k\text{mod}q)$ requires q computing operations, to evaluate it for all t_0 and $k\text{mod}q$ requires pq^2 operations. Note, the key is that less frequencies need to be evaluated when calculating $A(t_0, k\text{mod}q)$, in particular q frequencies rather than N. After evaluating $\{A(t_0, k_0); 0 \leq t_0 \leq p-1, 0 \leq k_0 \leq q-1\}$ we then need to take the Fourier transform of this over t_0 to evaluate $d(\omega_k)$ which is p operations and this needs to be done n times (to get all $\{d(\omega_k)\}_k$) this leads to np. Thus in total this leads to

$$\underbrace{p^2q}_{\text{evaluation of all A}} + \underbrace{np}_{\text{evaluation of the transforms of A}} = pq^2 + pn = n(q+p). \tag{A.14}$$

Observe that $n(p+q)$ is a lot smaller than n^2.

Looking back at the above calculation we observe that q^2 operations were required to calculate $A(t_0, k\text{mod}q) = A(t_0, k_0)$ for all $0 \leq k_0 \leq q-1$. However $A(t_0, k_0)$ is a Fourier transform

$$A(t_0, k_0) = \sum_{t_1=0}^{q-1} x_{t_1p+t_0} \exp\left[it_1\omega_{k_0,q}\right].$$

Therefore, we can use the same method as was used above to reduce this number. To do this we

need to factorize q into $p = p_1 q_1$ and using the above method we can write this as

$$\begin{aligned} A(t_0, k_0) &= \sum_{t_2=0}^{p_1-1} \sum_{t_3=0}^{q_1-1} x_{(t_2+t_3 p_1)p+t_0} \exp\left[i(t_2 + t_3 p_1)\omega_{k_0,q}\right] \\ &= \sum_{t_2=0}^{p_1-1} \exp\left[it_2\omega_{k_0,q}\right] \sum_{t_3=0}^{q_1-1} x_{(t_2+t_3 p_1)p+t_0} \exp\left[it_3 p_1 \omega_{k_0,q}\right] \\ &= \sum_{t_2=0}^{p_1-1} \exp\left[it_2\omega_{k_0,q}\right] \sum_{t_3=0}^{q_1-1} x_{(t_2+t_3 p_1)p+t_0} \exp\left[it_3\omega_{k_0 \mathrm{mod} q_1,q_1}\right]. \end{aligned}$$

We note that $k_0 \mathrm{mod} q_1 = (k \mathrm{mod}(p_1 q_1) \mathrm{mod} q_1) = k \mathrm{mod} q_1$, substituting this into the above we have

$$\begin{aligned} A(t_0, k_0) &= \sum_{t_2=0}^{p_1-1} \exp\left[it_2\omega_{k_0,q}\right] \sum_{t_3=0}^{q_1-1} x_{(t_2+t_3 p_1)p+t_0} \exp\left[it_3\omega_{k_0 \mathrm{mod} q_1,q_1}\right] \\ &= \sum_{t_2=0}^{p_1-1} \exp\left[it_2\omega_{k_0,q}\right] \underbrace{A(t_0, t_2, k_0 \mathrm{mod} q_1)}_{q_1 \text{ frequencies}}. \end{aligned}$$

Thus we see that q_1 computing operations are required to calculate $A(t_0, t_2, k_0 \mathrm{mod} q_1)$ and to calculate $A(t_0, t_2, k \mathrm{mod} q_1)$ for all $0 \leq t_2 \leq p_1 - 1$ and $0 \leq k \mathrm{mod} q_1 \leq q_1 - 1$ requires in total $q_1^2 p_1$ computing operations. After evaluating $\{A(t_0, t_2, k_0 \mathrm{mod} q_1); 0 \leq t_2 \leq q_2 - 1, 0 \leq k \mathrm{mod} q_1 \leq q_1 - 1\}$ we then need to take its Fourier transform over t_2 to evaluate $A(t_0, k_0)$, which is p_1 operations. Thus in total to evaluate $A(t_0, k_0)$ over all k_0 we require $q_1^2 p_1 + p_1 q$ operations. Thus we have reduced the number of computing operations for $A(t_0, k_0)$ from q^2 to $q(p_1 + q_1)$, substituting this into (A.14) gives the total number of computing operations to calculate $\{d(\omega_k)\}$

$$pq(p_1 + q_1) + np = n(p + p_1 + q_1).$$

In general the same idea can be used to show that given the prime factorization of $n = \prod_{s=1}^{m} p_s$, then the number of computing operations to calculate the DFT is $n(\sum_{s=1}^{m} p_s)$.

Example A.5.1 *Let us suppose that $n = 2^m$ then we can write $d(\omega_k)$ as*

$$\begin{aligned} d(\omega_k) = \sum_{t=1}^{n} x_t \exp(it\omega_k) &= \sum_{t=1}^{n/2} X_{2t} \exp(i2t\omega_k) + \sum_{t=0}^{(n/2)-1} X_{2t+1} \exp(i(2t+1)\omega_k) \\ &= \sum_{t=1}^{n/2} X_{2t} \exp(i2t\omega_k) + \exp(i\omega_k) \sum_{t=0}^{(n/2)-1} X_{2t+1} \exp(i2t\omega_k) \\ &= A(0, k\text{mod}(n/2)) + \exp(i\omega_k) A(1, k\text{mod}(n/2)), \end{aligned}$$

since $\sum_{t=1}^{n/2} X_{2t} \exp(i2t\omega_k)$ and $\sum_{t=1}^{n/2} X_{2t+1} \exp(i2t\omega_k)$ are the Fourier transforms of $\{X_t\}$ on a coarser scale, therefore we can only identify the frequencies on a coarser scale. It is clear from the above that the evaluation of $A(0, k\text{mod}(n/2))$ for $0 \leq k\text{mod}(n/2) \leq n/2$ requires $(n/2)^2$ operations and same for $A(1, k\text{mod}(n/2))$. Thus to evaluate both $A(0, k\text{mod}(n/2))$ and $A(1, k\text{mod}(n/2))$ requires $2(n/2)^2$ operations. Then taking the Fourier transform of these two terms over all $0 \leq k \leq n-1$ is an additional $2n$ operations leading to

$$2(n/2)^2 + 2n = n^2/2 + 2n \text{ operations } < n^2.$$

We can continue this argument and partition

$$\begin{aligned} A(0, k\text{mod}(n/2)) &= \sum_{t=1}^{n/2} X_{2t} \exp(i2t\omega_k) \\ &= \sum_{t=1}^{n/4} X_{4t} \exp(i4t\omega_k) + \exp(i2\omega_k) \sum_{t=0}^{(n/4)-1} X_{4t+2} \exp(i4t\omega_k). \end{aligned}$$

Using the same argument as above the calculation of this term over all k requires $2(n/4)^2 + 2(n/2) = n^2/8 + n$ operations. The same decomposition applies to $A(1, k\text{mod}(n/2))$. Thus calculation of both terms over all k requires $2[n^2/8 + n] = n^2/4 + 2n$ operations. In total this gives

$$(n^2/4 + 2n + 2n) \text{operations.}$$

Continuing this argument gives $mn = n \log_2 n$ operations, which is the often cited rate.

Typically, if the sample size is not of order 2^m zeros are added to the end of the sequence (called padding) to increase the length to 2^m.

Appendix B

Mixingales

In this section we prove some of the results stated in the previous sections using mixingales.

We first define a mixingale, noting that the definition we give is not the most general definition.

Definition B.0.1 (Mixingale) *Let $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots)$, $\{X_t\}$ is called a mixingale if it satisfies*

$$\rho_{t,k} = \left\{ \mathrm{E}\bigg(\mathrm{E}(X_t|\mathcal{F}_{t-k}) - \mathrm{E}(X_t) \bigg)^2 \right\}^{1/2},$$

where $\rho_{t,k} \to 0$ as $k \to \infty$. We note if $\{X_t\}$ is a stationary process then $\rho_{t,k} = \rho_k$.

Lemma B.0.1 *Suppose $\{X_t\}$ is a mixingale. Then $\{X_t\}$ almost surely satisfies the decomposition*

$$X_t = \sum_{j=0}^{\infty} \Big\{ \mathrm{E}(X_t|\mathcal{F}_{t-j-1}) - \mathrm{E}(X_t|\mathcal{F}_{t-j-1}) \Big\}. \tag{B.1}$$

PROOF. We first note that by using a telescoping argument that

$$X_t - \mathrm{E}(X_t) \;=\; \sum_{k=0}^{m} \big\{ \mathrm{E}(X_t|\mathcal{F}_{t-k}) - \mathrm{E}(X_t|\mathcal{F}_{t-k-1}) \big\} + \big\{ \mathrm{E}(X_t|\mathcal{F}_{t-m-1}) - \mathrm{E}(X_t) \big\}.$$

By definition of a martingale $\mathrm{E}\big(\mathrm{E}(X_t|\mathcal{F}_{t-m-1}) - \mathrm{E}(X_t)\big)^2 \to 0$ as $k \to \infty$, hence the remainder term in the above expansion becomes negligable as $m \to \infty$ and we have almost surely

$$\begin{aligned} & X_t - \mathrm{E}(X_t) \\ = \;& \sum_{k=0}^{\infty} \big\{ \mathrm{E}(X_t|\mathcal{F}_{t-k}) - \mathrm{E}(X_t|\mathcal{F}_{t-k-1}) \big\}. \end{aligned}$$

Thus giving the required result. □

We observe that (B.1) resembles the Wold decomposition. The difference is that the Wolds decomposition decomposes a stationary process into elements which are the errors in the best linear predictors. Whereas the result above decomposes a process into sums of martingale differences.

It can be shown that functions of several ARCH-type processes are mixingales (where $\rho_{t,k} \leq K\rho^k$ $(rho < 1)$), and Subba Rao (2006) and Dahlhaus and Subba Rao (2007) used these properties to obtain the rate of convergence for various types of ARCH parameter estimators. In a series of papers, Wei Biao Wu considered properties of a general class of stationary processes which satisfied Definition B.0.1, where $\sum_{k=1}^{\infty} \rho_k < \infty$.

In Section B.2 we use the mixingale property to prove Theorem 14.7.3. This is a simple illustration of how useful mixingales can be. In the following section we give a result on the rate of convergence of some random variables.

B.1 Obtaining almost sure rates of convergence for some sums

The following lemma is a simple variant on a result proved in Móricz (1976), Theorem 6.

Lemma B.1.1 *Let $\{\mathcal{S}_T\}$ be a random sequence where $\mathrm{E}(\sup_{1\leq t\leq T}|\mathcal{S}_t|^2) \leq \phi(T)$ and $\{phi(t)\}$ is a monotonically increasing sequence where $\phi(2^j)/\phi(2^{j-1}) \leq K < \infty$ for all j. Then we have almost surely*

$$\frac{1}{T}\mathcal{S}_T = O\Big(\frac{\sqrt{\phi(T)(\log T)(\log\log T)^{1+\delta}}}{T}\Big).$$

PROOF. The idea behind the proof is to that we find a subsequence of the natural numbers and define a random variables on this subsequence. This random variable, should 'dominate' (in some sense) S_T. We then obtain a rate of convergence for the subsequence (you will see that for the subsequence its quite easy by using the Borel-Cantelli lemma), which, due to the dominance, can be transfered over to S_T. We make this argument precise below.

Define the sequence $V_j = \sup_{t\leq 2^j}|S_t|$. Using Chebyshev's inequality we have

$$P(V_j > \varepsilon) \leq \frac{\phi(2^j)}{\varepsilon}.$$

Let $\varepsilon(t) = \sqrt{\phi(t)(\log\log t)^{1+\delta}\log t}$. It is clear that

$$\sum_{j=1}^{\infty} P(V_j > \varepsilon(2^j)) \leq \sum_{j=1}^{\infty} \frac{C\phi(2^j)}{\phi(2^j)(\log j)^{1+\delta} j} < \infty,$$

where C is a finite constant. Now by Borel Cantelli, this means that almost surely $V_j \leq \varepsilon(2^j)$. Let us now return to the orginal sequence S_T. Suppose $2^{j-1} \leq T \leq 2^j$, then by definition of V_j we have

$$\frac{S_T}{\varepsilon(T)} \leq \frac{V_j}{\varepsilon(2^{j-1})} \stackrel{a.s}{\leq} \frac{\varepsilon(2^j)}{\varepsilon(2^{j-1})} < \infty$$

under the stated assumptions. Therefore almost surely we have $S_T = O(\varepsilon(T))$, which gives us the required result. □

We observe that the above result resembles the law of iterated logarithms. The above result is very simple and nice way of obtaining an almost sure rate of convergence. The main problem is obtaining bounds for $\mathrm{E}(\sup_{1\leq t\leq T}|\mathcal{S}_t|^2)$. There is on exception to this, when $\mathcal{S}_t$ is the sum of martingale differences then one can simply apply Doob's inequality, where $\mathrm{E}(\sup_{1\leq t\leq T}|\mathcal{S}_t|^2) \leq \mathrm{E}(|\mathcal{S}_T|^2)$. In the case that S_T is not the sum of martingale differences then its not so straightforward. However if we can show that S_T is the sum of mixingales then with some modifications a bound for $\mathrm{E}(\sup_{1\leq t\leq T}|\mathcal{S}_t|^2)$ can be obtained. We will use this result in the section below.

B.2 Proof of Theorem 14.7.3

We summarise Theorem 14.7.3 below.

Theorem 1 *Let us suppose that $\{X_t\}$ has an ARMA representation where the roots of the characteristic polynomials $\phi(z)$ and $\theta(z)$ lie are greater than $1+\delta$. Then*

(i)

$$\frac{1}{n}\sum_{t=r+1}^{n} \varepsilon_t X_{t-r} = O(\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}}) \tag{B.2}$$

(ii)

$$\frac{1}{n}\sum_{t=\max(i,j)}^{n} X_{t-i}X_{t-j} = O(\sqrt{\frac{(\log\log n)^{1+\gamma}\log n}{n}}). \tag{B.3}$$

for any $\gamma > 0$.

By using Lemma **??**, and that $\sum_{t=r+1}^{n} \varepsilon_t X_{t-r}$ is the sum of martingale differences, we prove Theorem 14.7.3(i) below.

PROOF of Theorem 14.7.3. We first observe that $\{\varepsilon_t X_{t-r}\}$ are martingale differences, hence we can use Doob's inequality to give $\mathrm{E}(\sup_{r+1\leq s\leq T}(\sum_{t=r+1}^{s} \varepsilon_t X_{t-r})^2) \leq (T-r)\mathrm{E}(\varepsilon_t^2)\mathrm{E}(X_t^2)$. Now we can apply Lemma **??** to obtain the result. □

We now show that

$$\frac{1}{T}\sum_{t=\max(i,j)}^{T} X_{t-i}X_{t-j} = O(\sqrt{\frac{(\log\log T)^{1+\delta}\log T}{T}}).$$

However the proof is more complex, since $\{X_{t-i}X_{t-j}\}$ are not martingale differences and we cannot directly use Doob's inequality. However by showing that $\{X_{t-i}X_{t-j}\}$ is a mixingale we can still show the result.

To prove the result let $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots)$ and $\mathcal{G}_t = \sigma(X_{t-i}X_{t-j}, X_{t-1-i}X_{t-j-i}, \ldots)$. We observe that if $i > j$, then $\mathcal{G}_t \subset \mathcal{F}_{t-i}$.

Lemma B.2.1 *Let* $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots)$ *and suppose* X_t *comes from an ARMA process, where the roots are greater than* $1+\delta$. *Then if* $\mathrm{E}(\varepsilon_t^4) < \infty$ *we have*

$$\mathrm{E}\big(\mathrm{E}(X_{t-i}X_{t-j}|\mathcal{F}_{t-\min(i,j)-k}) - \mathrm{E}(X_{t-i}X_{t-j})\big)^2 \leq C\rho^k.$$

PROOF. By expanding X_t as an MA(∞) process we have

$$\begin{aligned} &\mathrm{E}(X_{t-i}X_{t-j}|\mathcal{F}_{t-\min(i,j)-k}) - \mathrm{E}(X_{t-i}X_{t-j}) \\ = &\sum_{j_1,j_2=0}^{\infty} a_{j_1}a_{j_2}\big\{\mathrm{E}(\varepsilon_{t-i-j_1}\varepsilon_{t-j-j_2}|\mathcal{F}_{t-k-\min(i,j)}) - \mathrm{E}(\varepsilon_{t-i-j_1}\varepsilon_{t-j-j_2})\big\}. \end{aligned}$$

Now in the case that $t-i-j_1 > t-k-\min(i,j)$ and $t-j-j_2 > t-k-\min(i,j)$, $\mathrm{E}(\varepsilon_{t-i-j_1}\varepsilon_{t-j-j_2}|\mathcal{F}_{t-k-\min(i,j)}) = \mathrm{E}(\varepsilon_{t-i-j_1}\varepsilon_{t-j-j_2})$. Now by considering when $t-i-j_1 \leq t-k-\min(i,j)$ or $t-j-j_2 \leq t-k-\min(i,j)$ we have have the result. □

Lemma B.2.2 *Suppose* $\{X_t\}$ *comes from an ARMA process. Then*

(i) The sequence $\{X_{t-i}X_{t-j}\}_t$ *satisfies the mixingale property*

$$\mathrm{E}\big(\mathrm{E}(X_{t-i}X_{t-j}|\mathcal{F}_{t-\min(i,j)-k}) - \mathrm{E}(X_{t-i}X_{t-j}|\mathcal{F}_{t-k-1})\big)^2 \leq K\rho^k, \tag{B.4}$$

and almost surely we can write $X_{t-i}X_{t-j}$ *as*

$$X_{t-i}X_{t-j} - \mathrm{E}(X_{t-i}X_{t-j}) = \sum_{k=0}^{\infty}\sum_{t=\min(i,j)}^{n} V_{t,k} \tag{B.5}$$

where $V_{t,k} = \mathrm{E}(X_{t-i}X_{t-j}|\mathcal{F}_{t-k-\min(i,j)}) - \mathrm{E}(X_{t-i}X_{t-j}|\mathcal{F}_{t-k-\min(i,j)-1})$*, are martingale differences.*

(ii) Furthermore $\mathrm{E}(V_{t,k}^2) \leq K\rho^k$ *and*

$$\mathrm{E}\Big\{\sup_{\min(i,j)\leq s\leq n}\Big(\sum_{t=\min(i,j)}^{s}\{X_{t-i}X_{t-j} - \mathrm{E}(X_{t-i}X_{t-j})\}\Big)^2\Big\} \leq Kn, \tag{B.6}$$

where K *is some finite constant.*

PROOF. To prove (i) we note that by using Lemma B.2.1 we have (B.4). To prove (B.5) we use the same telescoping argument used to prove Lemma B.0.1.

To prove (ii) we use the above expansion to give

$$\begin{aligned}
&\mathrm{E}\Big\{\sup_{\min(i,j)\leq s\leq n}\Big(\sum_{t=\min(i,j)}^{s}\{X_{t-i}X_{t-j} - \mathrm{E}(X_{t-i}X_{t-j})\}\Big)^2\Big\} \qquad \text{(B.7)}\\
=\ &\mathrm{E}\Big\{\sup_{\min(i,j)\leq s\leq n}\Big(\sum_{k=0}^{\infty}\sum_{t=\min(i,j)}^{s} V_{t,k}\Big)^2\Big\}\\
=\ &\mathrm{E}\Big\{\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\sup_{\min(i,j)\leq s\leq n}\Big|\sum_{t=\min(i,j)}^{s} V_{t,k_1}\Big| \times \Big|\sum_{t=\min(i,j)}^{s} V_{t,k_2}\Big|\Big\}\\
=\ &\Bigg(\sum_{k=0}^{\infty}\Bigg\{\mathrm{E}\Bigg(\sup_{\min(i,j)\leq s\leq n}\Big|\sum_{t=\min(i,j)}^{s} V_{t,k_1}\Big|^2\Bigg)\Bigg\}^{1/2}\Bigg)^2
\end{aligned}$$

Now we see that $\{V_{t,k}\}_t = \{\mathrm{E}(X_{t-i}X_{t-j}|\mathcal{F}_{t-k-\min(i,j)}) - \mathrm{E}(X_{t-i}X_{t-j}|\mathcal{F}_{t-k-\min(i,j)-1})\}_t$, therefore $\{V_{t,k}\}_t$ are also martingale differences. Hence we can apply Doob's inequality to $\mathrm{E}\big\{\sup_{\min(i,j)\leq s\leq n}\big(\sum_{t=\min(i,j)}^{s} V_{t,k}\big)$

and by using (B.4) we have

$$\mathrm{E}\Big\{\sup_{\min(i,j)\leq s\leq n}\Big(\sum_{t=\min(i,j)}^{s}V_{t,k}\Big)^2\Big\}\leq \mathrm{E}\Big(\sum_{t=\min(i,j)}^{n}V_{t,k}\Big)^2=\sum_{t=\min(i,j)}^{n}\mathrm{E}(V_{t,k}^2)\leq K\cdot n\rho^k.$$

Therefore now by using (B.7) we have

$$\mathrm{E}\Big\{\sup_{\min(i,j)\leq s\leq n}\Big(\sum_{t=\min(i,j)}^{s}\{X_{t-i}X_{t-j}-\mathrm{E}(X_{t-i}X_{t-j})\}\Big)^2\Big\}\leq Kn.$$

Thus giving (B.6). □

We now use the above to prove Theorem 14.7.3(ii).

PROOF of Theorem 14.7.3(ii). To prove the result we use (B.6) and Lemma B.1.1. □

Bibliography

Hong-Zhi An, Zhao-Guo. Chen, and E.J. Hannan. Autocorrelation, autoregression and autoregressive approximation. *Ann. Statist.*, 10:926–936, 1982.

R. B. Ash. *Real Analysis and Probability.* Academic Press, 1972.

A. Aue, L. Horvath, and J. Steinbach. Estimation in random coefficient autoregressive models. *Journal of Time Series Analysis*, 27:61–76, 2006.

K. I. Beltrao and P. Bloomfield. Determining the bandwidth of a kernel spectrum estimate. *Journal of Time Series Analysis*, 8:23–38, 1987.

I. Berkes, L. Horváth, and P. Kokoskza. GARCH processes: Structure and estimation. *Bernoulli*, 9:2001–2007, 2003.

I. Berkes, L. Horvath, P. Kokoszka, and Q. Shao. On discriminating between long range dependence and changes in mean. *Ann. Statist.*, 34:1140–1165, 2006.

R.N. Bhattacharya, V.K. Gupta, and E. Waymire. The hurst effect under trend. *J. Appl. Probab.*, 20:649–662, 1983.

P. Billingsley. *Probability and Measure.* Wiley, New York, 1995.

T Bollerslev. Generalized autoregressive conditional heteroscedasticity. *J. Econometrics*, 31:301–327, 1986.

P. Bougerol and N. Picard. Stationarity of GARCH processes and some nonnegative time series. *J. Econometrics*, 52:115–127, 1992a.

P. Bougerol and N Picard. Strict stationarity of generalised autoregressive processes. *Ann. Probab.*, 20:1714–1730, 1992b.

G. E. P. Box and G. M. Jenkins. *Time Series Analysis, Forecasting and Control.* Cambridge University Press, Oakland, 1970.

A. Brandt. The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Adv. in Appl. Probab.*, 18:211–220, 1986.

W.L. Briggs and V. E. Henson. *The DFT: An Owner's manual for the Discrete Fourier Transform.* SIAM, Philadelphia, 1997.

D.R. Brillinger. *Time Series: Data Analysis and Theory.* SIAM Classics, 2001.

P. Brockwell and R. Davis. *Time Series: Theory and Methods.* Springer, New York, 1998.

P. Brockwell and R. Davis. *Introduction to Time Series and Forecasting.* Springer, 2002.

R. Dahlhaus and D. Janas. A frequency domain bootstrap for ratio statistics in time series analysis. *Annals of Statistics*, 24:1934–1963, 1996.

R. Dahlhaus and S. Subba Rao. A recursive online algorithm for the estimation of time-varying arch parameters. *Bernoulli*, 13:389–422, 2007.

L. Dalla, V. Giraitis and P. C.B. Philips. Robust tels for white noise and cross correlation. *Preprint*, 2019.

J Davidson. *Stochastic Limit Theory.* Oxford University Press, Oxford, 1994.

Jérôme Dedecker and Paul Doukhan. A new covariance inequality and applications. *Stochastic Process. Appl.*, 106(1):63–80, 2003.

H. Dette and E. Paparoditis. Bootstrapping frequency domain tests in multivariate time series with an application to comparing spectral densities. *J. Royal Statistical Society (B)*, 71:831–857, 2009.

R. Douc, E. Moulines, and D. Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R Examples.* Chapman and Hall, 2014.

Y. Dwivedi and S. Subba Rao. A test for second order stationarity based on the discrete fourier transform. *Journal of Time Series Analysis*, 32:68–91, 2011.

R. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica*, 50:987–1006, 1982.

Philip A. Ernst and Paul Shaman. The bias mapping of the Yule-Walker estimator is a contraction. *Statist. Sinica*, 29:1831–1849, 2019.

J. C. Escanciano and I. N Lobato. An automatic Portmanteau test for serial correlation. *Journal of Econometrics*, 151:140–149, 2009.

J. Fan and Q. Yao. *Nonlinear time series: Nonparametric and parametric methods*. Springer, Berlin, 2003.

J. Franke and W. Härdle. On bootstrapping kernel spectral estimates. *Ann. Statist.*, 20:121–145, 1992.

W. Fuller. *Introduction to Statistical Time Series*. Wiley, New York, 1995.

C. W. J. Granger and A. P. Andersen. *An introduction to Bilinear Time Series models*. Vandenhoek and Ruprecht, Göttingen, 1978.

U. Grenander and G. Szegö. *Toeplitz forms and Their applications*. Univ. California Press, Berkeley, 1958.

P Hall and C.C. Heyde. *Martingale Limit Theory and its Application*. Academic Press, New York, 1980.

E.J. Hannan and Rissanen. Recursive estimation of ARMA order. *Biometrika*, 69:81–94, 1982.

J. Hart. Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society*, 53:173–187, 1991.

C. Hurvich and S. Zeger. Frequency domain bootstrap methods for time series. *New York University, Graduate School of Business Adminstration*, 1987.

C. Jentsch and S. Subba Rao. A test for second order stationarity of multivariate time series. *Journal of Econometrics*, 2014.

D. A. Jones. Nonlinear autoregressive processes. *Proceedings of the Royal Society (A)*, 360:71–95, 1978.

R. H. Jones. Maximum likelihood fitting of arma models to time series with missing observations. *Technometrics*, 22:389–395, 1980.

J. Krampe, J-P. Kreiss, and Paparoditis. Estimated wold representation and spectral density driven bootstrap for time series. *Technical Report*, 2016.

J.-P. Kreiss. *Bootstrapping and Related Techniques*, chapter Bootstrap procedures for AR(∞)-processes, pages 107–113. Springer, 1992.

Jens-Peter Kreiss, Efstathios Paparoditis, and Dimitris N. Politis. On the range of validity of the autoregressive sieve bootstrap. *Ann. Statist.*, 39, 2011.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17(3):1217–1241, 1989.

S. N. Lahiri. *Resampling methods for dependent data*. Springer Series in Statistics. Springer-Verlag, New York, 2003.

W. K. Li. On the asymptotic standard errors of residual autocorrelations in nonlinear time series modelling. *Biometrika*, 79:435–437, 1992.

Regina Y. Liu and Kesar Singh. Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the limits of bootstrap (East Lansing, MI, 1990)*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pages 225–248. Wiley, New York, 1992.

I. N. Lobato. Testing that a dependent process is uncorrelated. *Journal of the American Statistical Association*, 96:1066–1076, 2001.

H. Lütkepohl. *A new introduction to multiple time series analysis*. Springer, Berlin, 2005.

T. Mikosch. *Elementary Stochastic Calculus With Finance in View*. World Scientific, 1999.

T. Mikosch and C. Stărică. Is it really long memory we see in financial returns? In P. Embrechts, editor, *Extremes and Integrated Risk Management*, pages 149–168. Risk Books, London, 2000.

T. Mikosch and C. Stărică. Long-range dependence effects and arch modelling. In P. Doukhan, G. Oppenheim, and M.S. Taqqu, editors, *Theory and Applications of Long Range Dependence*, pages 439–459. Birkhäuser, Boston, 2003.

F. Móricz. Moment inequalities and the strong law of large numbers. *Z. Wahrsch. verw. Gebiete*, 35:298–314, 1976.

D.F. Nicholls and B.G. Quinn. *Random Coefficient Autoregressive Models, An Introduction.* Springer-Verlag, New York, 1982.

E. Parzen. On consistent estimates of the spectrum of a stationary process. *Ann. Math. Statist.*, 1957.

E. Parzen. On estimation of the probability density function and the mode. *Ann. Math. Statist.*, 1962.

D.N. Politis and J. P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313, 1994.

M. Pourahmadi. *Foundations of Time Series Analysis and Prediction Theory.* Wiley, 2001.

M. B. Priestley. *Spectral Analysis and Time Series: Volumes I and II.* Academic Press, London, 1983.

B.G. Quinn and E.J. Hannan. *The Estimation and Tracking of Frequency.* Cambridge University Press, 2001.

M. Rosenblatt and U. Grenander. *Statistical Analysis of Stationary Time Series.* Chelsea Publishing Co, 1997.

Paul Shaman and Robert A. Stine. The bias of autoregressive coefficient estimators. *J. Amer. Statist. Assoc.*, 83(403):842–848, 1988.

X. Shao. A self-normalized approacj to confidence interval construction in time series. *Journal of the Royal Statistical Society (B)*, 72:343–366, 2010.

Xiaofeng Shao and Wei Biao Wu. Local Whittle estimation of fractional integration for nonlinear processes. *Econometric Theory*, 23(5):899–929, 2007.

R. Shumway and D. Stoffer. *Time Series Analysis and Its applications: With R examples.* Springer, New York, 2006.

V. Statulevicius and Jakimavicius. Estimates of semiivariant and centered moments of stochastic processes with mixing: I. *Lithuanian Math. J.*, 28:226–238, 1988.

D. Straumann. *Estimation in Conditionally Heteroscedastic Time Series Models*. Springer, Berlin, 2005.

S. Subba Rao. A note on uniform convergence of an arch(∞) estimator. *Sankhya*, pages 600–620, 2006.

S. Subba Rao and J. Yang. Reconciling the Gaussian and Whittle likelihood with an application to estimation in the frequency domain. *arXiv preprint arXiv:2001.06966*, 2020.

T. Subba Rao. On the estimation of bilinear time series models. In *Bull. Inst. Internat. Statist. (paper presented at 41st session of ISI, New Delhi, India)*, volume 41, 1977.

T. Subba Rao. On the theory of bilinear time series models. *Journal of the Royal Statistical Society(B)*, 43:244–255, 1981.

T. Subba Rao and M. M. Gabr. A test for linearity of a stationary time series. *J of Time Series Analysis*, 1:145–158, 1980.

T. Subba Rao and M. M. Gabr. *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Lecture Notes in Statistics (24). Springer, New York, 1984.

S. C. Taylor. *Modelling Financial Time Series*. John Wiley and Sons, Chichester, 1986.

Gy. Terdik. *Bilinear Stochastic Models and Related Problems of Nonlinear Time Series Analysis; A Frequency Domain Approach*, volume 142 of *Lecture Notes in Statistics*. Springer Verlag, New York, 1999.

M. Vogt. Nonparametric regression for locally stationary time series. *Annals of Statistics*, 40: 2601–2633, 2013.

A. M. Walker. On the estimation of a harmonic component in a time series with stationary independent residuals. *Biometrika*, 58:21–36, 1971.

P. Whittle. Gaussian estimation in stationary time series. *Bulletin of the International Statistical Institute*, 39:105–129, 1962.