A course in Time Series Analysis

Suhasini Subba Rao

Email: suhasini.subbarao@stat.tamu.edu

August 29, 2022

Contents

Preface

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

- You are encouraged to use `R` (see David Stoffer's tutorial). I have tried to include Rcode in the notes so that you can replicate some of the results.
- Exercise questions will be in the notes and will be set at regular intervals.
- Finally, these notes are dedicated to my wonderful Father, whose inquisitive questions, and unconditional support inspired my quest in time series.

Chapter 1

Introduction

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

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

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

1.1 Time Series data

Below we discuss four different data sets.

The Southern Oscillation Index from 1876-present

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

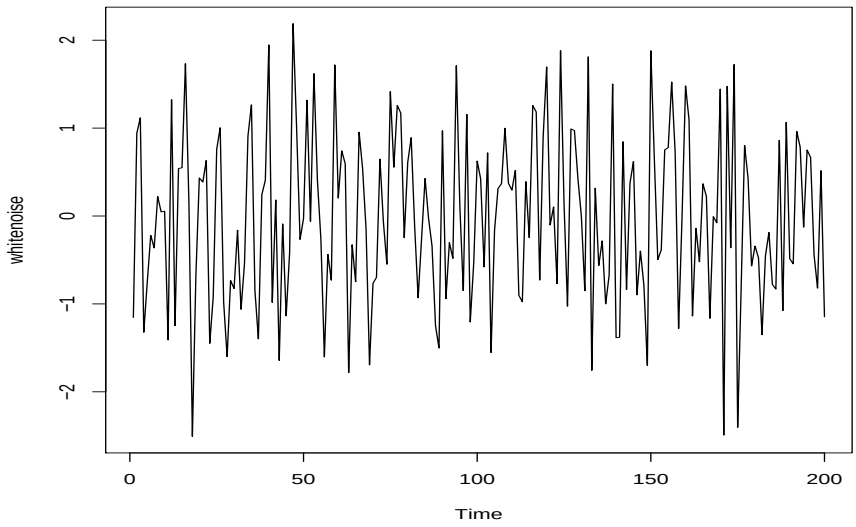


Figure 1.1: Plot of independent uncorrelated random variables

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

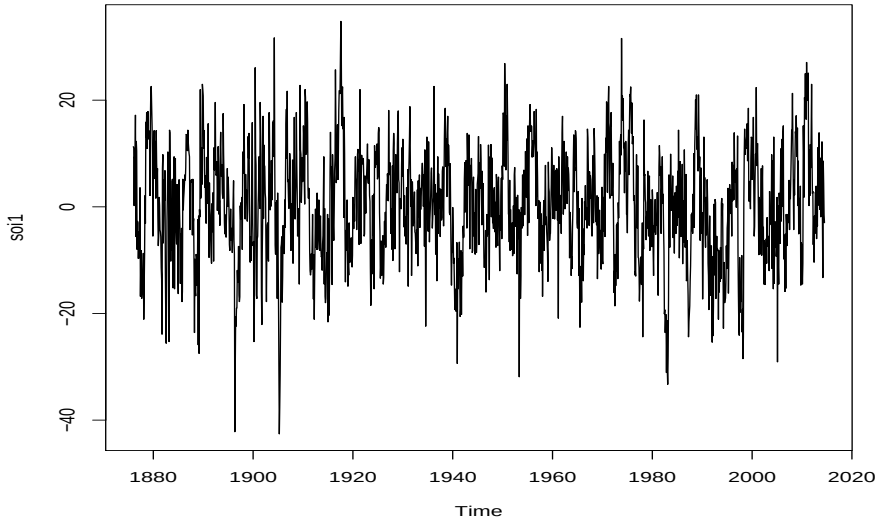


Figure 1.2: Plot of monthly Southern Oscillation Index, 1876-2014

Nasdaq Data from 1985-present

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

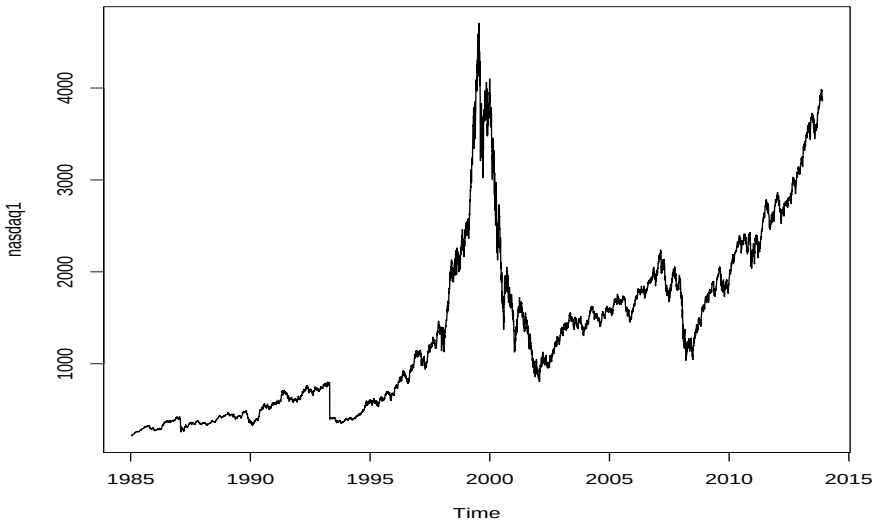


Figure 1.3: Plot of daily closing price of Nasdaq 1985-2014

Yearly sunspot data from 1700-2013

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

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

Yearly and monthly average temperature data

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

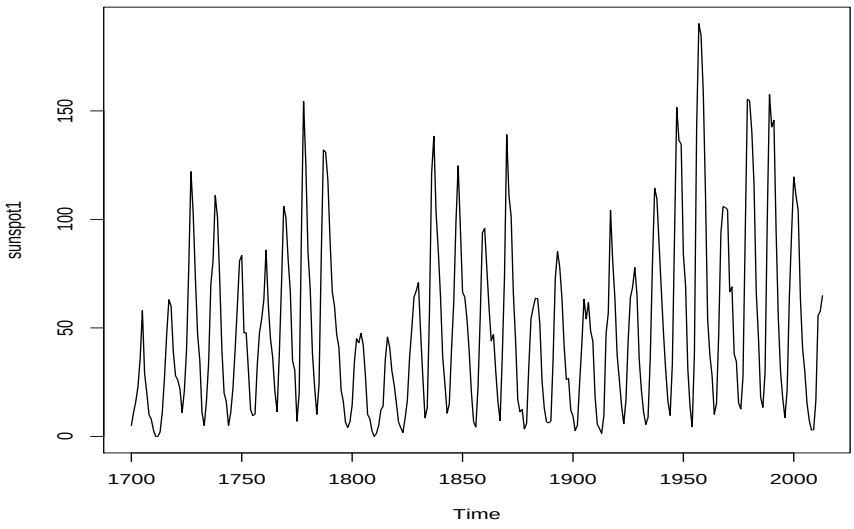


Figure 1.4: Plot of Sunspot numbers 1700-2013

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

1.2 R code

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

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

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

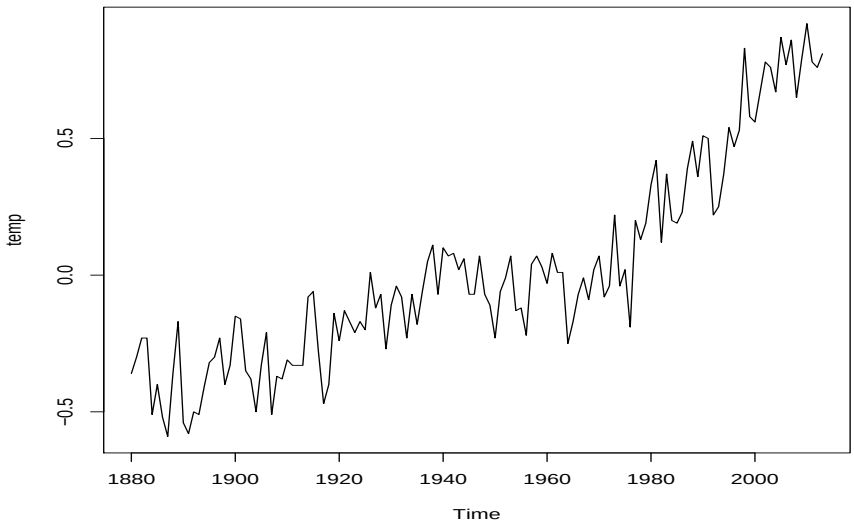


Figure 1.5: Plot of global, yearly average, temperature anomalies, 1880 - 2013

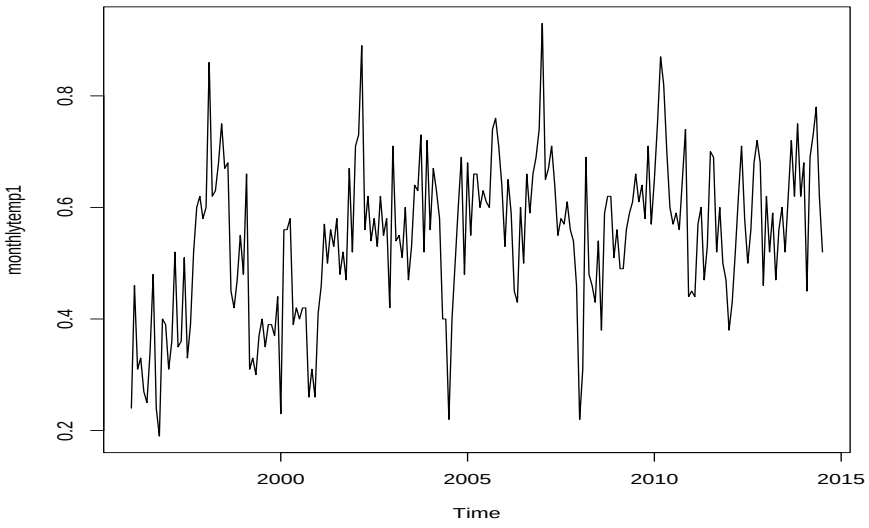


Figure 1.6: Plot of global, monthly average, temperatures January, 1996 - July, 2014.

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

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

1.3 Filtering time series

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

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

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

where $\{h_j\}$ are weights.

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

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

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

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

1.4 Terminology

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

Chapter 2

Trends in a time series

Objectives:

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

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

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

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

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

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

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

2.1 Parametric trend

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

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

or

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

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

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

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

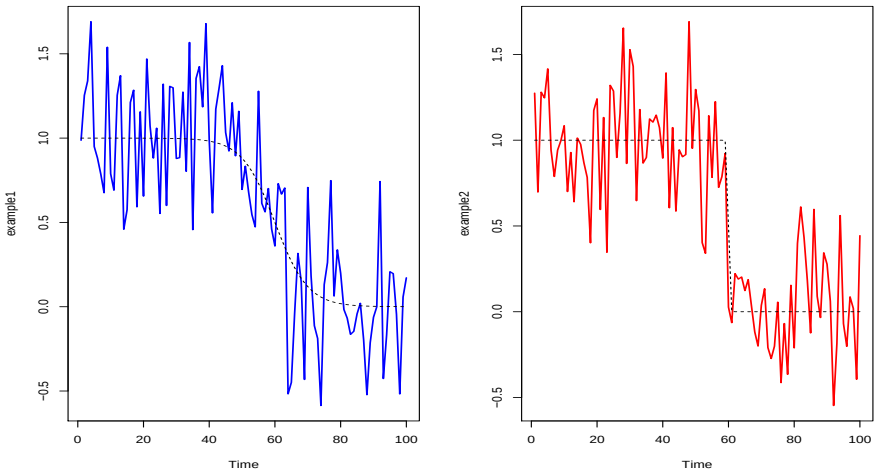


Figure 2.1: The function Y_t in (2.3) with iid noise with $\sigma = 0.3$. Dashed is the truth. Left: $\beta_0 = 0.2$ and $\beta_1 = 60$. Right: $\beta_0 = 5$ and $\beta_1 = 60$

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

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

Both these nonlinear parametric models motivate the general nonlinear model

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

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

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

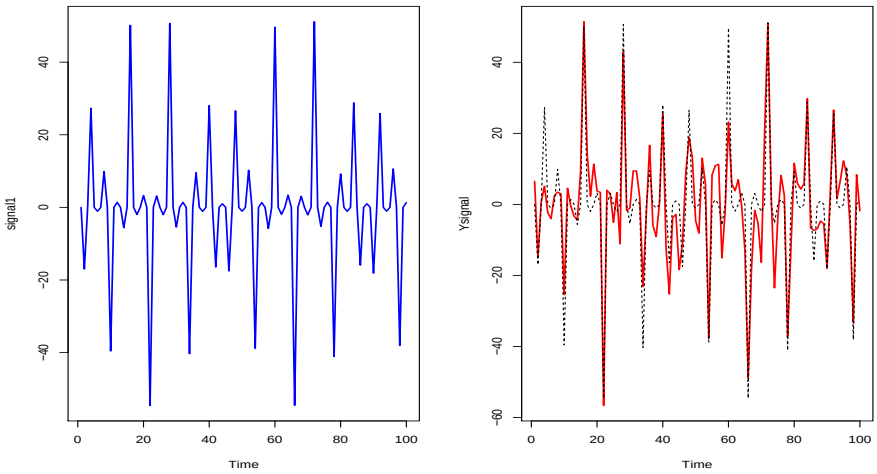


Figure 2.2: The Burst signal (equation (2.4)) $A = 1$, $\beta_0 = 2$, $\beta_1 = 1$ and $\theta = \pi/2$ with iid noise with $\sigma = 8$. Dashed is the truth. Left: True Signal. Right: True Signal with noise

with sines and cosines

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

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

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

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* $Y_i =$

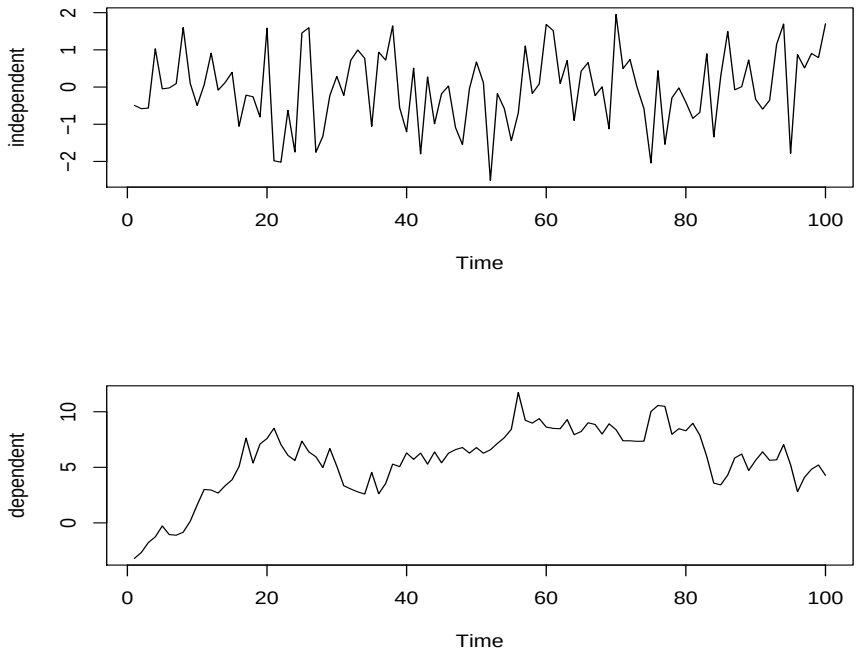


Figure 2.3: Top: realisations from iid random noise. Bottom: Realisation from dependent noise

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

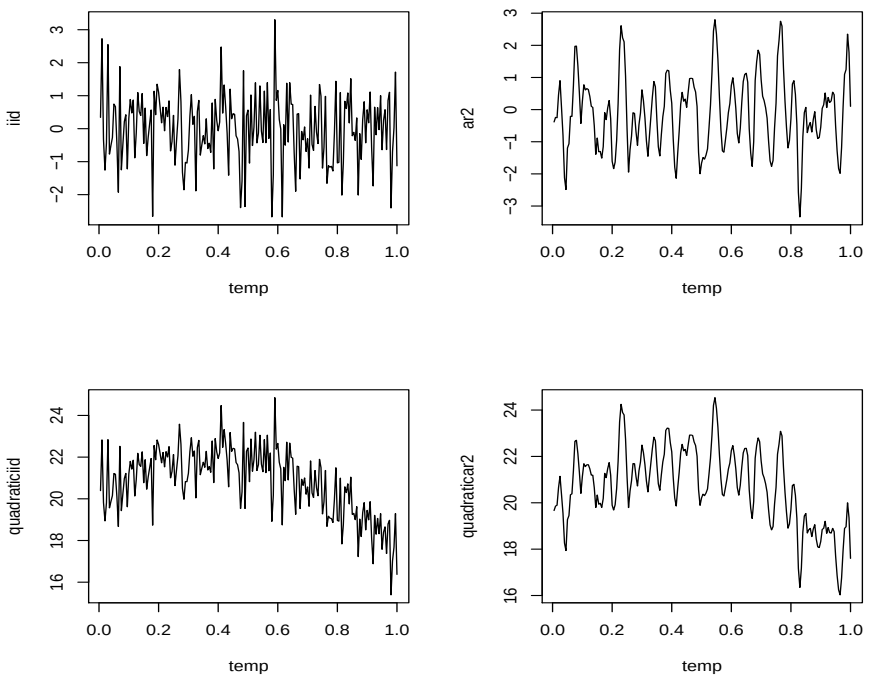


Figure 2.4: Top: realisations from iid random noise and dependent noise (left = iid and right = dependent). Bottom: Quadratic trend plus corresponding noise.

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 $d_P(t) =$

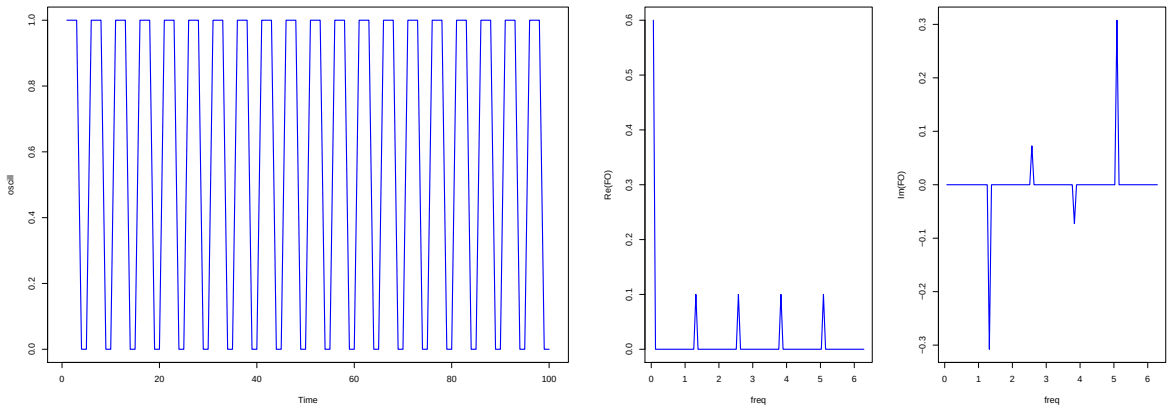


Figure 2.5: Left: Periodic function $d_5(s) = 1$ for $s = 1, 2$, $d_5(s) = 0$ for $s = 3, 4, 5$ (period 5), Right: The real and imaginary parts of its Fourier transform

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

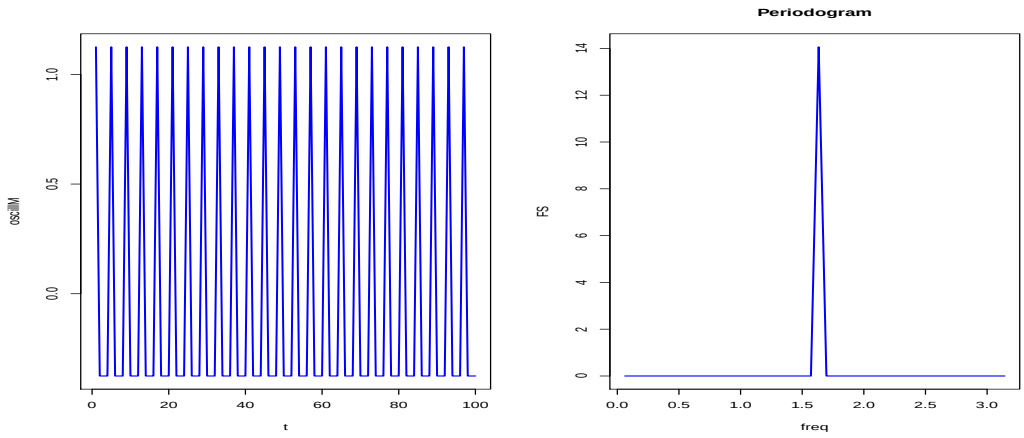


Figure 2.6: Left: Periodic function $d_4(s) = (1.125, -0.375, -0.375, -0.375)$ (period 4)

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.

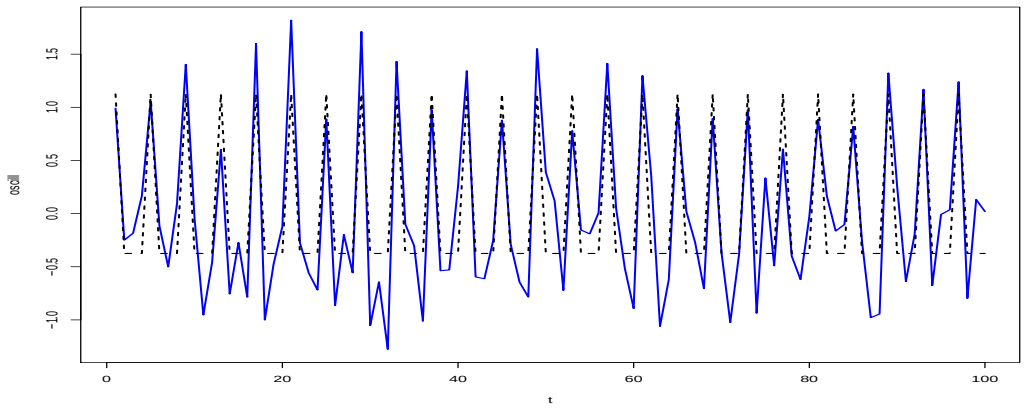


Figure 2.7: Periodic function $d_4(s) = (1.125, -0.375, -0.375, -0.375)$ (period 4) and signal with noise (blue line).

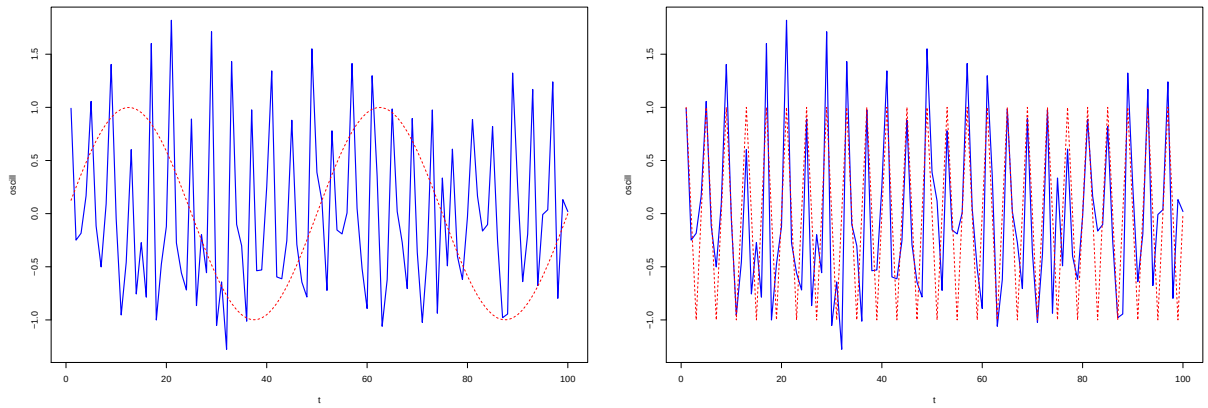


Figure 2.8: Left: Signal superimposed with $\sin(u)$. Right: Signal superimposed with $\sin(25u)$.

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

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

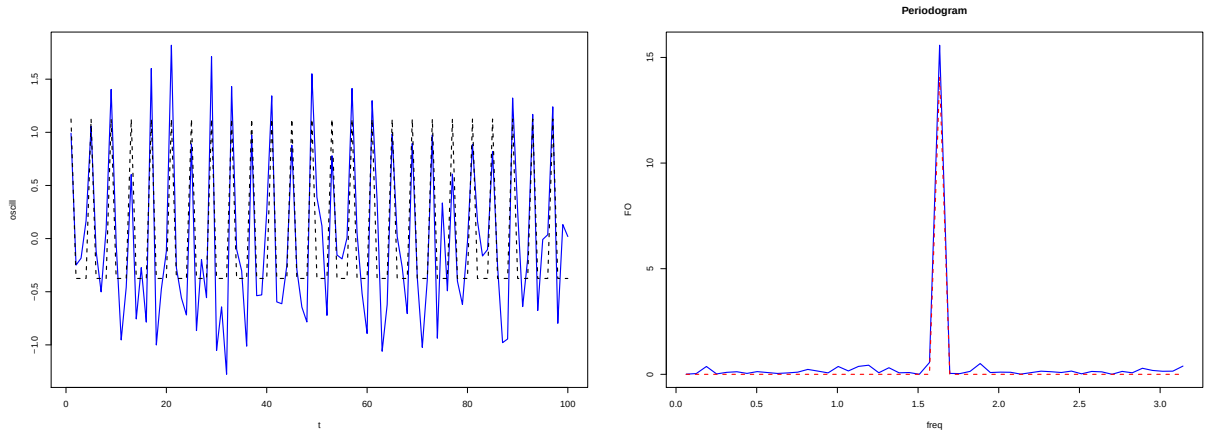


Figure 2.9: Left: Signal, Right: periodogram of signal (peridogram of periodic function in red)

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).

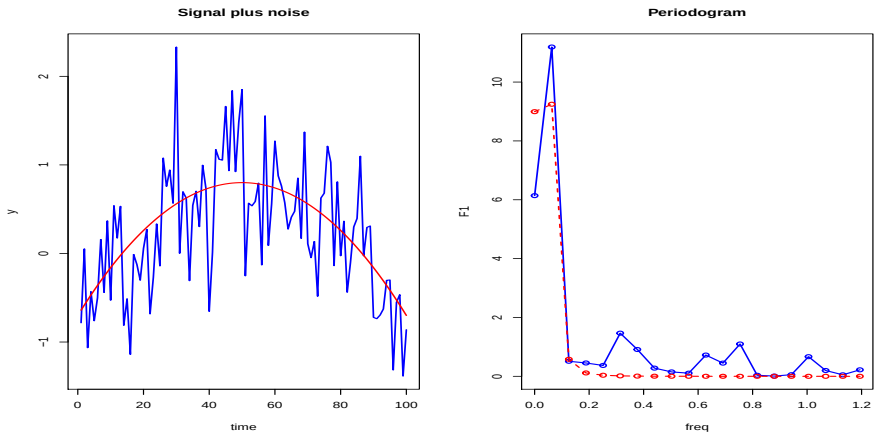


Figure 2.10: Left: Signal and noise (blue). The signal is in red. Right: Periodogram of signal plus noise (up to frequency $\pi/5$). Periodogram of signal is in red.

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*).*

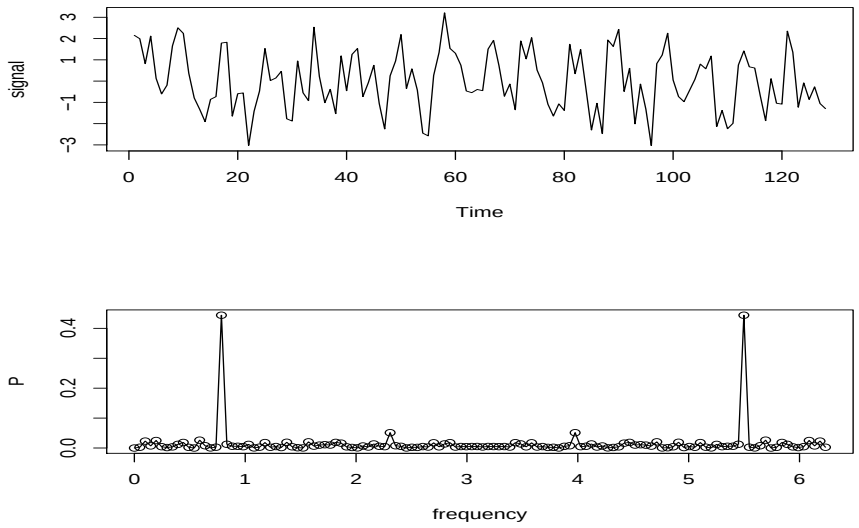


Figure 2.11: Left: Realisation of (2.18) plus iid noise, Right: Periodogram of signal plus iid noise.

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

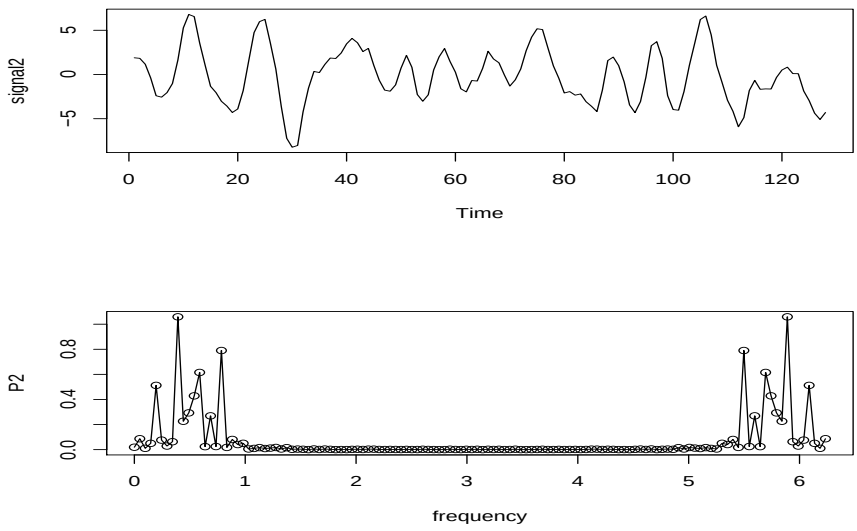


Figure 2.12: Top: Realisation of (2.18) plus correlated noise and $n = 128$, Bottom: Periodogram of signal plus correlated noise.

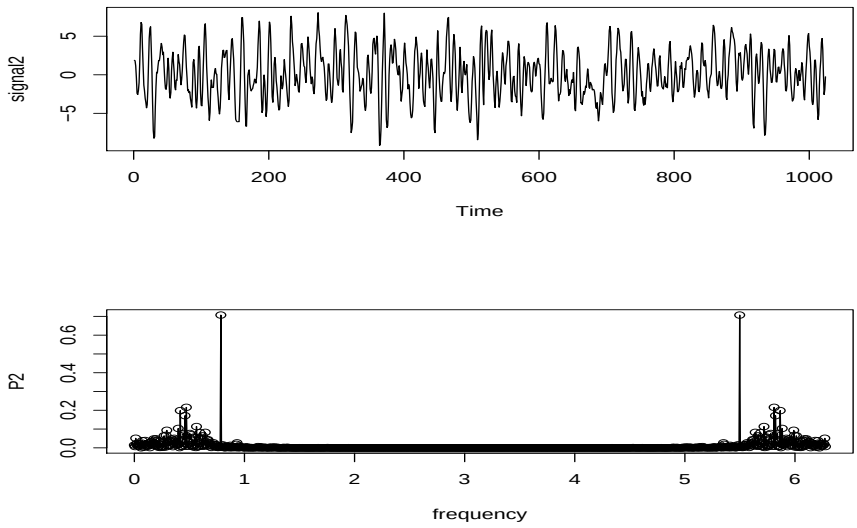


Figure 2.13: Top: Realisation of (2.18) plus correlated noise and $n = 1024$, Bottom: Periodogram of signal plus correlated noise.

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

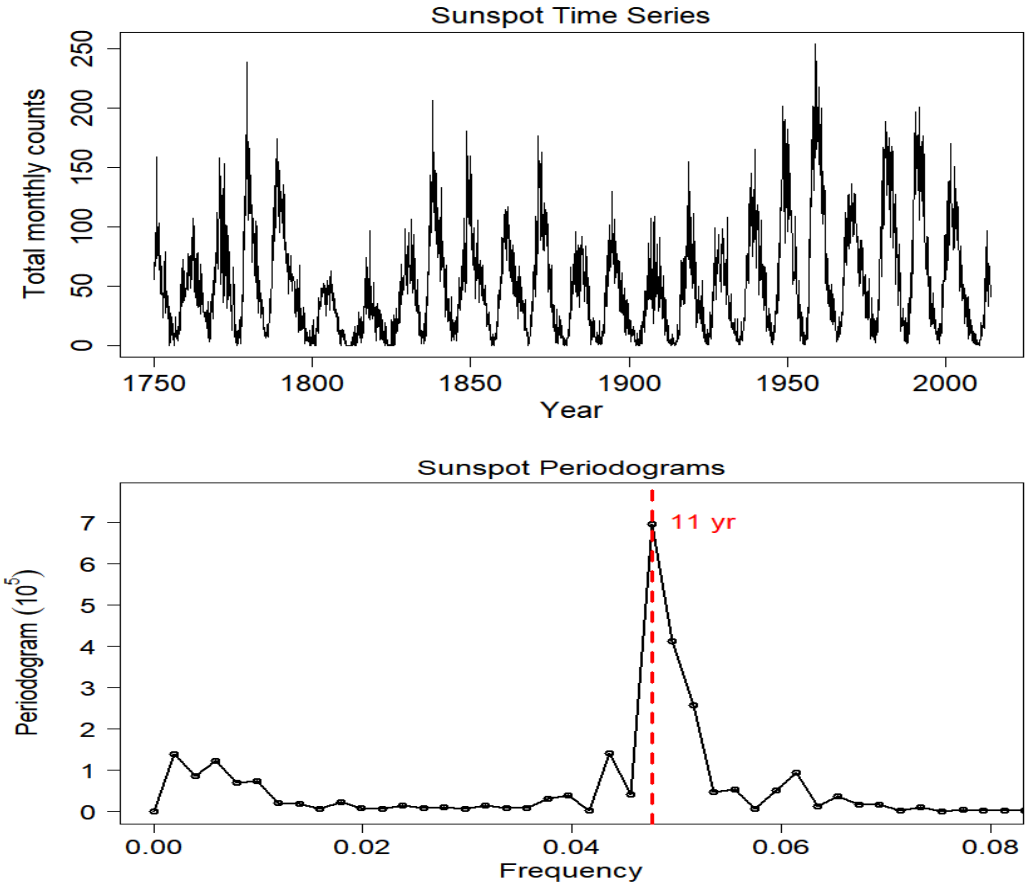


Figure 2.14: Sunspot data from Jan, 1749 to Dec, 2014. There is a peak at about 30 along the line which corresponds to $2\pi/P = 0.047$ and $P \approx 132$ months (11 years).

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

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

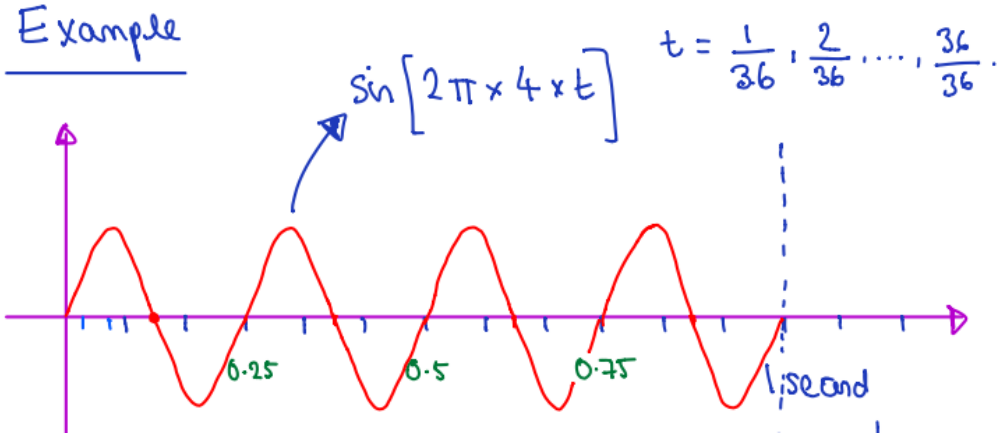
However, in the 1920's, Udny Yule, a statistician, and Gilbert Walker, a Meterologist (working in Pune, India) believed an alternative model could be used to explain the features seen in the periodogram. We consider their proposed approach in Section 4.3.5.

2.6 Data Analysis: EEG data

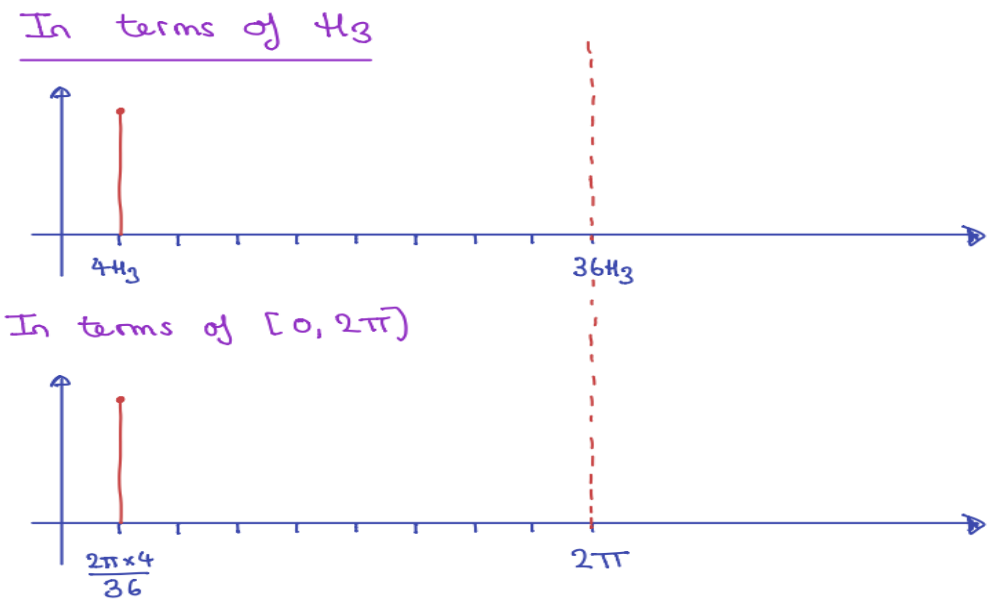
2.6.1 Connecting Hertz and Frequencies

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.



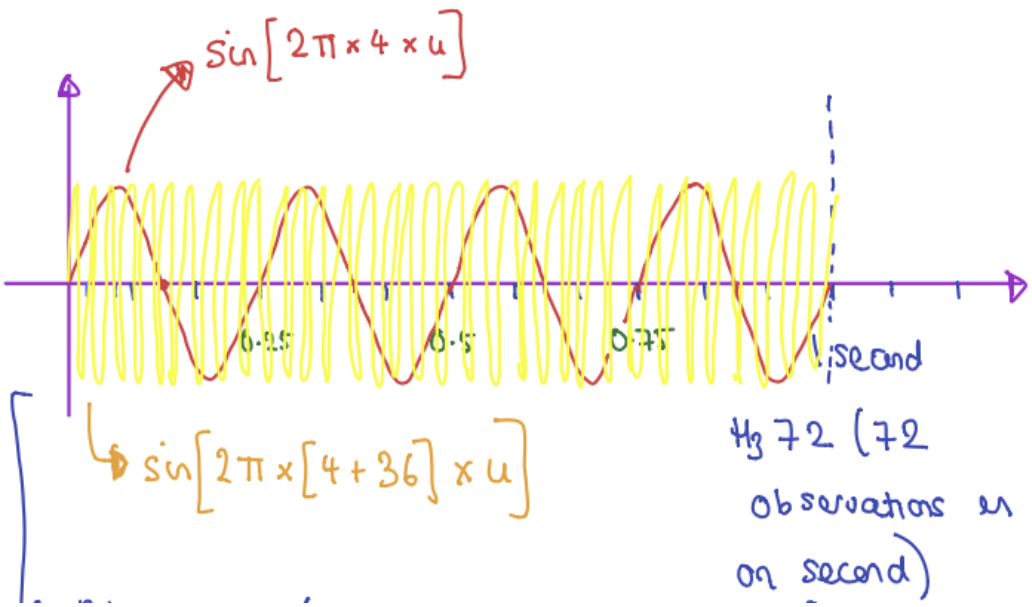
Analysis from the perspective of time series Typically, in time series, the sampling frequency is kept the same. Just the same number of second that the time series is observed grows. This allows us obtain a finer frequency grid on $[0, 2\pi]$ and obtain a better resolution in terms of peaks in frequencies.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

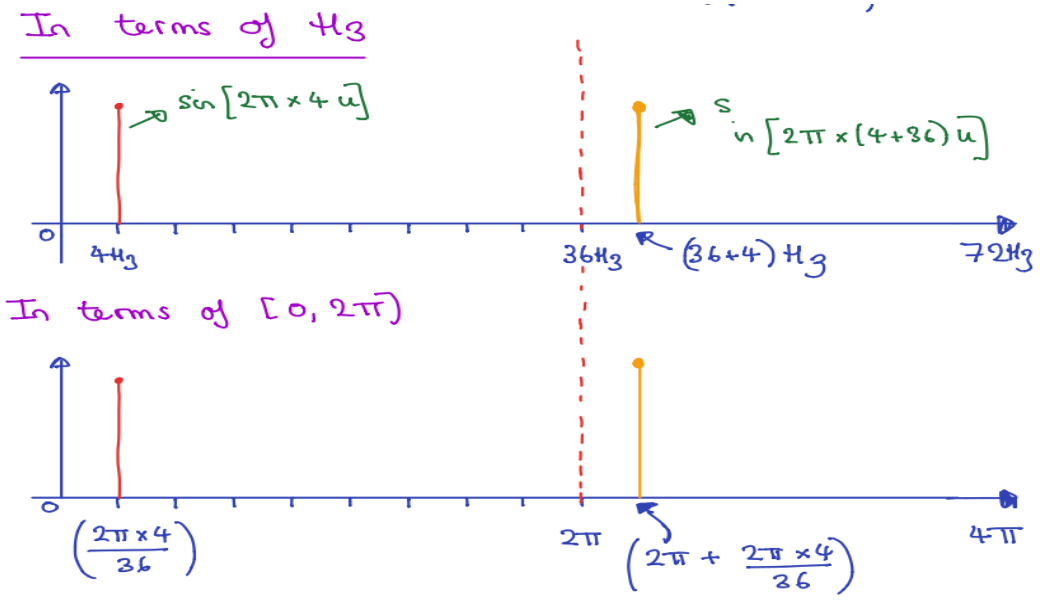
$$\sin\left(2\pi \times 4 \times \frac{t}{36}\right) = \sin\left(2\pi \times (4 + 36) \times \frac{t}{36}\right) \text{ for all } t \in \mathbb{Z}$$

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).

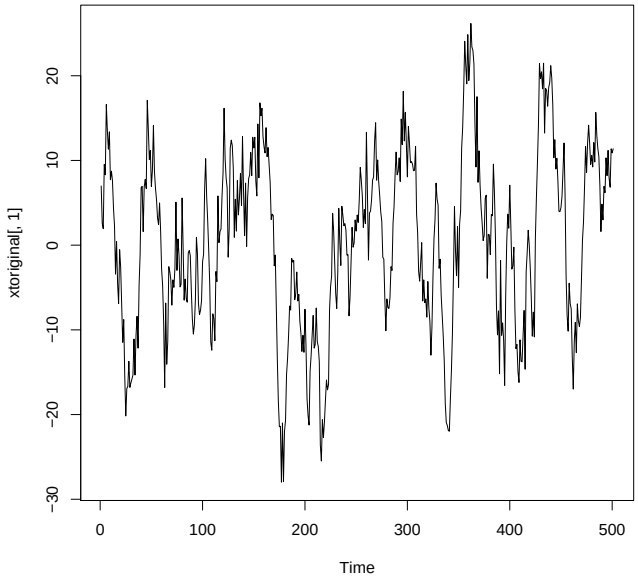


Figure 2.15: Original EEG..

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*

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

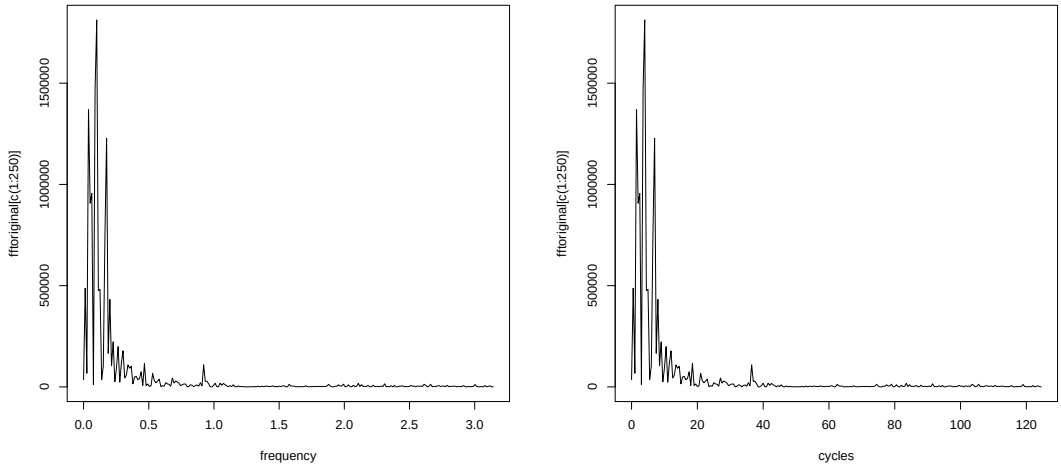


Figure 2.16: Left: Periodogram of original EEG on $[0, 2\pi]$. Right: Periodogram in terms of cycles per second.

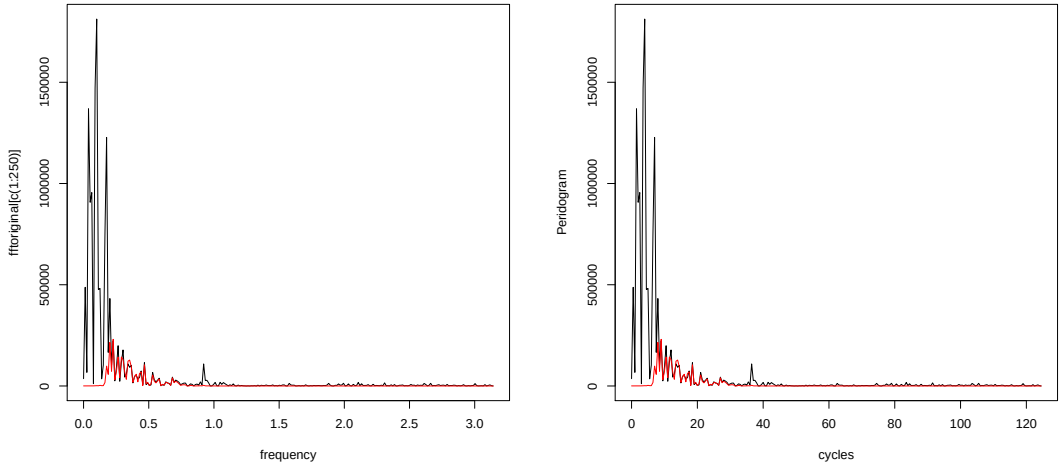


Figure 2.17: The periodogram of original EEG overlayed with processed EEG (in red). The same plot is given below, but the x-axis corresponds to cycles per second (measured in Hz)

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 $\{Y_t\}$.*

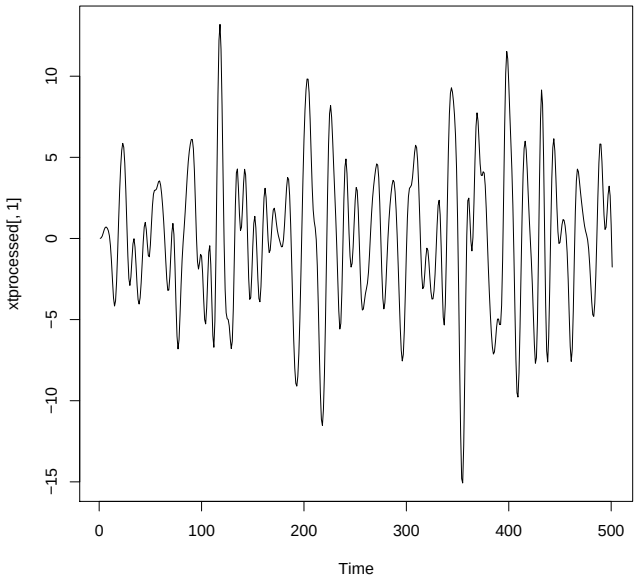


Figure 2.18: Time series after processing with a Buttersworth filter.

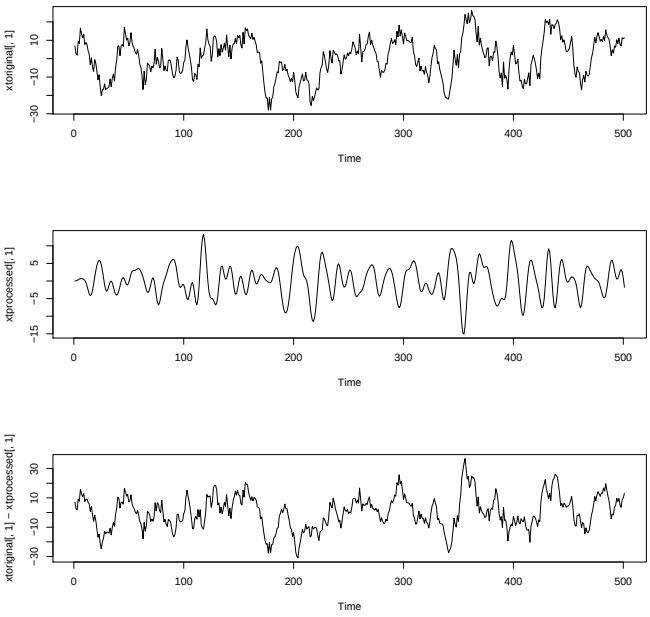


Figure 2.19: Top: Original EEG. Middle: Filtered EEG and Bottom: Difference between Original and Filtered EEG

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`*).*

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

2.7 Exercises

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

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

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

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

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

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

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

(i) Let

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

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* $Y_t=2\sin(2pit/8)+\varepsilon_t$ *where*

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

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

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

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

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

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

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

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

R Code

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

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

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

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

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

Chapter 3

Stationary Time Series

3.1 Preliminaries

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

Different types of convergence

(i) Almost sure convergence: $X_n \stackrel{\text{a.s.}}{\to} a$ 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.

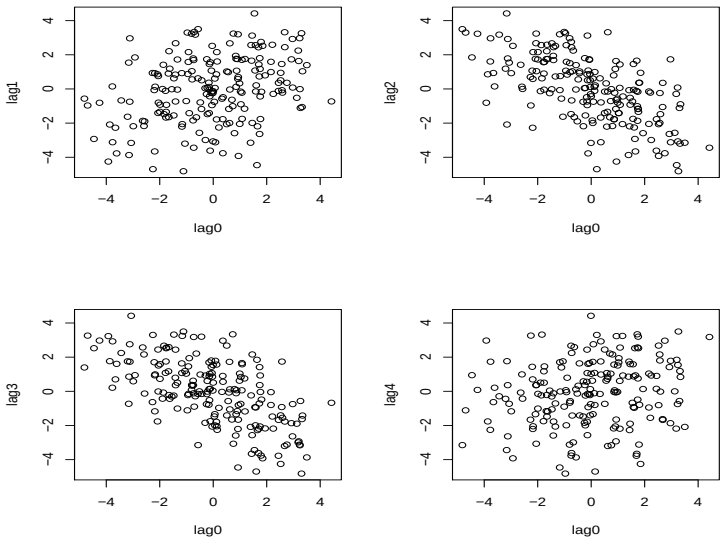


Figure 3.1: From model (3.3). Plot of X_t against X_{t+r} for $r = 1, \ldots, 4$. Top left: $r = 1$. Top right: $r = 2$, Bottom left: $r = 3$ and Bottom right: $r = 4$.

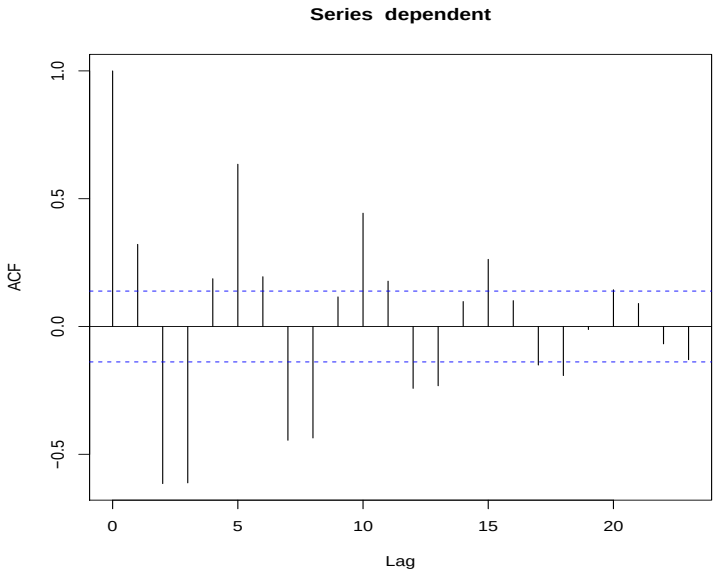


Figure 3.2: ACF plot of realisation from model (3.3).

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 $\operatorname{var}\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right] = O(n^{-1})$. Estimation of the variance of $\widehat{\boldsymbol{\beta}}_n$ is important and requires one to estimate $\operatorname{var}\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{u}_t \varepsilon_t\right)$. This is often done using the HAC estimator. We describe how this is done in Section 8.5.

3.3 Stationary processes

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

were given independent random variables all with different distributions, what parameter would you be estimating, it is not possible to estimate anything!).

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

3.3.1 Types of stationarity

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

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

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

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

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

Remark 3.3.1 (Strict and second order stationarity) *(i) If a process is strictly stationarity 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 not 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* $\mathrm{E}[V_t^2] < \infty$*.*

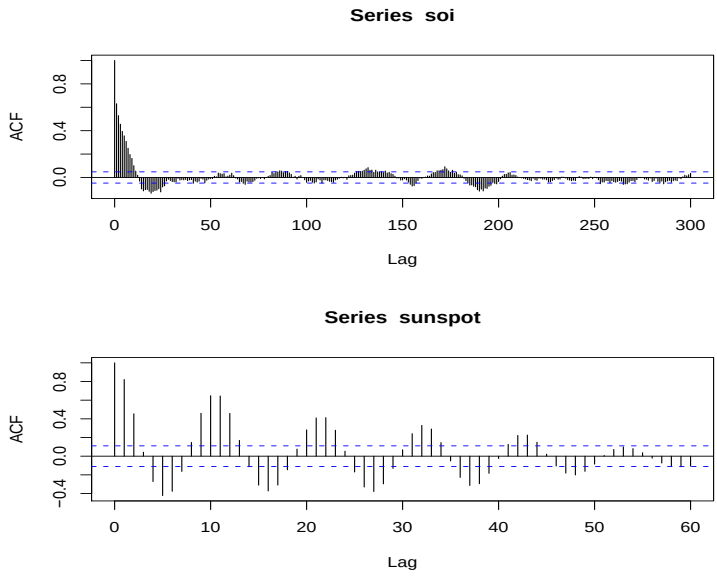


Figure 3.3: Top: ACF of Southern Oscillation data. Bottom ACF plot of Sunspot data.

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

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