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.
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](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.
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](image)

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 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
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.

1.1.1 R code

A large number of the methods and concepts will be illustrated in R. If you are not familiar with this language please learn the very 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,
# where 1 denotes January. Frequency = number of observations in a
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.

# unit of time (year). As the data is monthly it is 12.
plot.ts(soil)

Dating plots properly is very useful. This can be done using the package zoo and the function as.Date.
1.2 Detrending a time series

In time series, the main focus is on modelling the relationship between observations. Time series analysis is usually performed after the data has been detrended. In other words, if $Y_t = \mu_t + \varepsilon_t$, where $\{\varepsilon_t\}$ is zero mean time series, we first estimate $\mu_t$ and then conduct the time series analysis on the residuals. Once the analysis has been performed, we return to the trend estimators and use the results from the time series analysis to construct confidence intervals etc. In this course the main focus will be on the data after detrending. However, we start by reviewing some well known detrending methods. A very good primer is given in Shumway and Stoffer, Chapter 2, and you are strongly encouraged to read it.

1.2.1 Parametric trend

Often a parametric trend is assumed. Common examples include a linear trend

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

and the quadratic trend

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t.$$  

(1.2)

For example we may fit such models to the yearly average temperature data. Alternatively we may want to include seasonal terms

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

For example, we may believe that the Southern Oscillation Index has a period 12 (since the observations are taken monthly) and we use sine and cosine functions to model the seasonality. For these type of models, least squares can be used to estimate the parameters.
1.2.2 Taking differences to avoid fitting linear and higher order trends

Let us return to the Nasdaq data. Here we see a different type of “trend” behaviour. This is often referred to as stochastic trend. For most financial data the stochastic trend is removed by taking first difference (after taking logarithms). First differencing also avoids the need of fitting a linear trend to a model. For example if \( Y_t = \beta_0 + \beta_1 t + \epsilon_t \), then

\[
Z_t = Y_{t+1} - Y_t = \beta_1 + \epsilon_{t+1} - \epsilon_t.
\]

Taking higher order differences (ie. taking first differences of \( \{Z_t\} \) removes quadratic terms) removes higher order polynomials. The number of differences corresponding to the order of the polynomial.

Beware, taking too many differences can induce “ugly” dependences in the data. Eg. If \( X_t \) are iid random variables then \( Z_t = X_t - X_{t-1} \) is a dependent random variable. So one should not over difference the data.

**Exercise 1.1**

(i) Import the yearly temperature data (file global_mean_temp.txt) into R and fit the linear model in (1.1) to the data (use the R command `lsfit`).

(ii) Suppose the errors in (1.1) are correlated. Under the correlated assumption, explain why the standard errors reported in the R output are unreliable. Actually they are not reported in the output! But they are usually calculated as

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

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

Make a plot of the first differences of the temperature data.

What do you notice about the two plots, similar?

The AIC (Akaike Information Criterion) is usually used to select the parameters in the model (see wiki). You should have studied the AIC/AICc/BIC in several of the prerequisites you
have taken. In this course it will be assumed that you are familiar with it.

1.2.3 Estimation using nonparametric methods

In Section 1.2.1 we assumed that the mean had a certain known parametric form. This may not always be the case. If we have no apriori idea of what features may be in the mean, we can estimate the mean trend using a nonparametric approach. If we do not have any apriori knowledge of the mean function we cannot estimate it without placing some assumptions on it’s structure. The most common is to assume that the mean $\mu_t$ is a sample from a ‘smooth’ function, ie. $\mu_t = \mu(\frac{x}{n})$. Under this assumption the following approaches are valid.

Possibly one of the most 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

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

where $0 < \lambda < 1$. The choice of $\lambda$ depends on how much weight one wants to give the present observation. It is straightforward to show that

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

where $\gamma = -\log(1 - \lambda)$. Let $b = 1/\gamma$ and $K(u) = \exp(-u)I(u \geq 0)$, then $\hat{\mu}_t$ can be written as

$$\hat{\mu}_t = (1 - e^{1/b}) \sum_{j=1}^{n} K\left(\frac{t - j}{b}\right) Y_j,$$

This we observe that the exponential rolling window estimator is very close to a nonparametric kernel estimator of the mean, which has the form

$$\tilde{\mu}_t = \sum_{j=1}^{n} \frac{1}{b} K\left(\frac{t - j}{b}\right) Y_j.$$
it is likely you came across such estimators in your nonparametric classes. 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 nonparametric kernel estimators we can be observations on both sides of the neighbourhood of \( t \).

Other type of estimators include sieve-estimators. This is where we expand \( \mu(u) \) in terms of an orthogonal basis \( \{ \phi_k(u); k \in \mathbb{Z} \} \)

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

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’ \( \phi_k \). Thus 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

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

The orthogonality of the basis means that the least squares estimator \( \hat{a}_k \) is

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

It is worth pointing out that regardless of the method used, correlations in the errors \( \{ \varepsilon_t \} \) will play an role in quality of the estimator and even on the choice of bandwidth, \( b \), or equivalently the number of basis functions, \( M \) (see Hart (1991)). To understand why, suppose the mean function is \( \mu_t = \mu \left( \frac{t}{200} \right) \) (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 (1.6)). The plots are given in Figure 1.7. 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.

![Graphs of noise and quadratic trend](image)

Figure 1.7: Top: realisations from iid random noise and dependent noise (left = iid and right = dependent). Bottom: Quadratic trend plus corresponding noise.

**Exercise 1.2** The purpose of this exercise is to understand the correlated errors in a non-parametric model influences local smoothing estimators. Define the smooth signal \( f(u) = 5 \cdot (2u - 2.5u^2) + 20 \) and suppose we observe \( Y_i = f(i/200) + \varepsilon_i \) (\( n = 200 \)). To simulate \( f(u) \) with \( n = 200 \) define \( \text{temp} <- c(1:200)/200 \) and \( \text{quadratic} <- 5 \cdot (2 \cdot \text{temp} - 2.5 \cdot (\text{temp}^2)) + 20 \).

(i) Simulate from the above model using iid noise. You can use the code \( \text{iid} = \text{rnorm}(200) \) and \( \text{quadraticiid} = (\text{quadratic} + \text{iid}) \).

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

(ii) Simulate from the above model using correlated noise (we simulate from an AR(2)) \( \text{ar2} = 0.5 \cdot \text{arima.sim}(\text{list(order=c(2,0,0), ar = c(1.5, -0.75))}, n=200) \) and define \( \text{quadraticar2} = (\text{quadratic} + \text{ar2}) \).
Again estimate \( f \) using local averages.

By making plots of you estimators against \texttt{temp} compare them.

### 1.2.4 Estimation of the period

Suppose that the observations \( \{Y_t; t = 1, \ldots, n\} \) satisfy the following regression model

\[
Y_t = A \cos(\omega t) + B \sin(\omega t) + \varepsilon_t
\]

where \( \{\varepsilon_t\} \) are iid standard normal random variables and \( 0 < \omega < \pi \). The parameters \( A, B, \) and \( \omega \) are real and unknown. Unlike the regression models given in (1.2.1) the model here is nonlinear, since the unknown parameter, \( \omega \), is inside a trigonometric function. Standard least squares methods cannot be used to estimate the parameters. Assuming Gaussianity of \( \{\varepsilon_t\} \), the maximum likelihood corresponding to the model is

\[
\mathcal{L}_n(A, B, \omega) = -\frac{1}{2} \sum_{t=1}^{n} (Y_t - A \cos(\omega t) - B \sin(\omega t))^2.
\]

Nonlinear least squares method (which would require the use of a numerical maximisation scheme) can be employed to estimate the parameters. However, using some algebraic manipulations, explicit expressions for the estimators can be obtained (see Walker (1971) and Exercise 1.4). These are

\[
\hat{\omega}_n = \arg \max_{\omega} I_n(\omega)
\]

where

\[
I_n(\omega) = \frac{1}{n} \left| \sum_{t=1}^{n} Y_t \exp(it\omega) \right|^2 = \frac{1}{n} \left( \sum_{t=1}^{n} Y_t \cos(t\omega) \right)^2 + \frac{1}{n} \left( \sum_{t=1}^{n} Y_t \sin(t\omega) \right)^2. \tag{1.4}
\]

(we look for the maximum over the fundamental frequencies \( \omega_k = \frac{2\pi k}{n} \) for \( 1 \leq k \leq n \)),

\[
\hat{A}_n = \frac{2}{n} \sum_{t=1}^{n} Y_t \cos(\hat{\omega}_n t) \quad \text{and} \quad \hat{B}_n = \frac{2}{n} \sum_{t=1}^{n} Y_t \sin(\hat{\omega}_n t).
\]
The rather remarkable aspect of this result is that the rate of convergence of $|\hat{\omega}_n - \omega| = O(n^{-3/2})$, which is faster than the standard $O(n^{-1/2})$ that we usually encounter (we will see this in Example 1.2.1).

**Remark 1.2.1** $I_n(\omega)$ is usually called the periodogram. We observe that $I_n(\omega) = |J_n(\omega)|^2$ where

$$J_n(\omega) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Y_t \exp(it\omega) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Y_t [\cos(t\omega) + i \sin(t\omega)] \quad i = \sqrt{-1}$$

In practice, we only evaluate $J_n(\omega)$ and $I_n(\omega)$ at the so called fundamental frequencies $\omega_k = \frac{2\pi k}{n}$.

$$\{Y_t\}_{t=1}^{n} \rightarrow \{J_n(\frac{2\pi k}{n})\}_{k=1}^{n}.$$ So $J_n(\omega_k)$ 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, it so happens that for stationary time series this so called Fourier transform has some advantages.

Searching for peaks in the periodogram is a long established method for detecting periodicities. If we believe that there were two or more periods in the time series, we can generalize the method to searching for the largest and second largest peak etc. We consider an example below.

**Example 1.2.1** Consider the following model

$$Y_t = 2\sin\left(\frac{2\pi t}{8}\right) + \varepsilon_t \quad t = 1, \ldots, n. \quad (1.5)$$

where $\varepsilon_t$ are iid standard normal random variables. 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 (1.4)). In Figure 1.8 we give a plot of one realisation together with a plot of the periodogram. We observe that there is a symmetry, this is because of the $\exp(\omega)$ in the definition of $I(\omega)$ we can show that $I(\omega) = I(2\pi - \omega)$. Notice there is a clear peak at frequency $2\pi/8 \approx 0.78$ (where we recall that 8 is the period).
This method works extremely well if the error process \( \{ \varepsilon_t \} \) is uncorrelated. However, problems arise when the errors are correlated. To illustrate this issue, consider again model (1.5) but this time let us suppose the errors are correlated. More precisely, they satisfy the AR(2) model,

\[
\varepsilon_t = 1.5 \varepsilon_{t-1} - 0.75 \varepsilon_{t-2} + \epsilon_t,
\]

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 2). As in the iid case we use a sample size \( n = 128 \). In Figure 1.9 we give a plot of one realisation and the corresponding periodogram. We observe that the peak at \( 2\pi/8 \) is not the highest. The correlated errors (often called coloured noise) is masking the peak by introducing new peaks. To see what happens for larger sample sizes, we consider exactly the same model (1.5) with the noise generated as in (1.6). 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 1.10. In contrast to the smaller sample size, a large peak is seen 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 disentangle the deterministic period from the noise. Indeed we will show in Chapters
Figure 1.9: Left: Realisation of (1.5) with correlated noise and $n = 128$, Right: Periodogram

Figure 1.10: Left: Realisation of (1.5) with correlated noise and $n = 1024$, Right: Periodogram

2 and 3 that linear time series can exhibit similar types of behaviour to a periodic deterministic signal. This is a subject of ongoing research that dates back at least 60 years (see Quinn and Hannan (2001)).

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.

(ii) The periodogram holds important properties about the correlations in the noise (observe the periodogram in both Figures 1.9 and 1.10), there is some interesting activity in the lower frequencies, that appear to be due to noise.

This is called spectral analysis and is explored in Chapters 8 and 9. Indeed a lot of time series analysis can be done within the so called frequency or time domain.

1.2.5 Historic Background

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 is given in Figure 1.11. The periodogram below shows a peak at $2 \times 30\pi/314$, which corresponds to a period of $314/30 \approx 10.5$ years. Assuming that the the sunspot data roughly follows the period trend plus noise model

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

this suggests that the number of suns follow a periodic cycle with a peak every 10.5 years.

The general view until the 1920s is 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. Yule fitted an Autoregressive model of order two to the Sunspot data and obtained the AR(2) model

$$X_t = 1.381X_{t-1} - 0.6807X_{t-2} + \varepsilon_t,$$
Figure 1.11: Sunspot data from 1700 to 2014. There is a peak at about 30 along the line which corresponds to $2 \times 30\pi/314$ and $314/30 \approx 10.5$ years.
this corresponds to the characteristic function $1 - 1.381x + 0.68x^2$, whose roots are $0.77^{-1} \exp(\pm i0.57)$. Yule showed that if the roots of the characteristic polynomial of an AR(2) process were complex (not on the unit circle), then the solution would have so called ‘pseudo periodicities’. The above model has the solution
\[
X_t = \sum_{j=0}^{\infty} 0.77^j \sin [0.57(j + 1)] \varepsilon_{t-j},
\]
(we learn why later on in this the course). We see that the solution is completely stochastic (no deterministic mean), but the sin/cos functions make a typical realisation ‘look’ periodic (though there is no real periodic). Thus giving the peaks in the corresponding periodogram. In Figure 1.12 we compare a periodogram of the sunspot data and a realisation from the fitted AR(2) process. In Figure 1.13 we make a plot of the sunspot data and a realisation of the AR(2) process.

1.2.6 Exercises

Exercise 1.3 (Understanding Fourier transforms)  
(i) Let $Y_t = 1$. Plot the Periodogram of \{\text{\{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 \{\text{\{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 \{\text{\{Y_t; t = 1, \ldots, 128\}}\}.

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

Exercise 1.4  
(i) Let
\[
S_n(A, B, \omega) = \left( \sum_{t=1}^{n} Y_t^2 - \sum_{t=1}^{n} Y_t(A \cos(\omega t) + B \sin(\omega t)) + \frac{1}{2}n(A^2 + B^2) \right).
\]
Figure 1.12: The periodogram of the Sunspot data is the top plot and the periodogram of the fitted AR(2) model is the lower plot. They do not look exactly the same, but the AR(2) model is able to capture some of the periodicities.
Figure 1.13: Top: Sunspot, Lower: a realisation from the AR(2) process. Lines correspond to period of $P = 2\pi / \tan^{-1}(0.903/1.381) = 10.85$ years.
Show that

\[
2\mathcal{L}_n(A, B, \omega) + S_n(A, B, \omega) = \frac{(A^2 - B^2)}{2} \sum_{t=1}^{n} \cos(2t\omega) - AB \sum_{t=1}^{n} \sin(2t\omega).
\]

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

Since \(\mathcal{L}_n(A, B, \omega)\) and \(-\frac{1}{2}S_n(A, B, \omega)\) are asymptotically equivalent (i) shows that we can maximise \(-\frac{1}{2}S_n(A, B, \omega)\) 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 \(\hat{\omega}_n = \arg\max_{\omega} |\sum_{t=1}^{n} Y_t \exp(it\omega)|^2\).

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

\[
\sum_{t=1}^{n} \exp(i\Omega t) = \left\{
\begin{array}{ll}
\frac{\exp(\frac{i}{2}(n+1)\Omega)\sin(\frac{1}{2}n\Omega)}{\sin(\frac{1}{2}\Omega)} & 0 < \Omega < 2\pi \\
n & \Omega = 0 \text{ or } 2\pi.
\end{array}
\right.
\]

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

\[
2I(A, B, \omega) = E\left(\frac{\partial^2 S_n}{\partial \omega^2}\right) = \begin{pmatrix}
\frac{n}{2} & 0 & \frac{n^2}{2}B + O(n) \\
0 & \frac{n}{2} & -\frac{n^2}{2}A + O(n) \\
\frac{n^2}{2}B + O(n) & -\frac{n^2}{2}A + O(n) & \frac{n^3}{3}(A^2 + B^2) + O(n^2)
\end{pmatrix}.
\]

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

Exercise 1.5 (i) Simulate three hundred times from model (1.5) using \(n = 128\). Esti-
mate \( \omega, A \) and \( B \) for each simulation and obtain the empirical mean squared error

\[
\frac{1}{300} \sum_{i=1}^{300} (\hat{\theta}_i - \theta)^2
\]

(where \( \theta \) denotes the parameter and \( \hat{\theta}_i \) the estimate).

In your simulations, is the estimate of the period, \( \omega \) superior to the estimator of coefficients, \( A \) and \( B \)?

(ii) Do the same as above but now use coloured noise given in (1.6) as the errors. How do your estimates compare to (i)?

R Code

Simulation and periodogram for model (1.5) with iid errors:

```r
temp <- rnorm(128)
signal <- 1.5*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")
```

Simulation and periodogram for model (1.5) with correlated errors:

```r
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")
```
1.3 Some formalism

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

(i) \( \Omega \) 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 \( \omega \), \( \{X_t(\omega)\} \) is not random. In time series we will usually assume that what we observe \( x_t = X_t(\omega) \) (for some \( \omega \)) 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 \( \Omega \) (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.

Different types of convergence we will be using in class:

(i) Almost sure convergence: \( X_n \xrightarrow{a.s.} a \) as \( n \to \infty \) (in this course \( a \) will always be a constant).

This means for every \( \omega \in \Omega \) \( X_n \to a \), where \( P(\Omega) = 1 \) (this is classical limit of a sequence, see Wiki for a definition).

(ii) Convergence in probability: \( X_n \xrightarrow{P} a \). 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 \( X_n \xrightarrow{2} a \). This means \( E|X_n - a|^2 \to 0 \) 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).

- Which implies which?

  - (i), (ii) and (iii) imply (iv).
  - (i) implies (ii).
– (iii) implies (ii).

• Central limit theorems require (iv). It is often easiest to show (iii) (since this only requires mean and variance calculations).

1.4 Estimating the mean

Based on one realisation of a time series we want to make inference about parameters associated with the process \{X_t\}, such as the mean etc. Let us consider the simplest case, estimating the mean. We recall that in classical statistics we usually assume we observe several independent realisations, \{X_t\} from a random variable \(X\), and use the multiple realisations to make inference about the mean: \(\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k\). Roughly speaking, by using several independent realisations we are sampling over the entire probability space and obtaining a good estimate of the mean. On the other hand if the samples were highly dependent, then it is likely that \{X_t\} would be concentrated over small parts of the probability space. In this case, the variance of the sample mean would not converge to zero as the sample size grows.

A typical time series is a half way house between totally dependent data and independent data. Unlike, classical statistics, in time series, parameter estimation is based on only one 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 estimates, of the mean, can be made, based on just one realisation 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 trajectory, 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 measure of linear dependency is the covariance. The covariance in the stochastic process \{X_t\} is defined as

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

Noting that if \(\{X_t\}\) has zero mean, then the above reduces to \(\text{cov}(X_t, X_{t+k}) = E(X_tX_{t+k})\).
Remark 1.4.1 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 occasions 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.

Returning to the sample mean example suppose that $\{X_t\}$ is a time series. In order to estimate the mean we need to be sure that the mean is constant over time (else the estimator will be meaningless). Therefore we will assume that $\{X_t\}$ is a time series with constant mean $\mu$. We observe $\{X_t\}_{t=1}^n$ and estimate the mean $\bar{X} = \frac{1}{n} \sum_{t=1}^n X_t$. It is clear that this is an unbiased estimator of $\mu$, since $\text{E}(\bar{X}) = \mu$ (it is unbiased). Thus to see whether it converges in mean square to $\mu$ we consider its variance

$$\text{var}(\bar{X}) = \frac{1}{n^2} \sum_{t=1}^n \text{var}(X_t) + \frac{2}{n^2} \sum_{t=1}^n \sum_{\tau=t+1}^n \text{cov}(X_t, X_\tau). \tag{1.8}$$

If the covariance structure decays at such a rate that the sum of all lags is finite (sup$_t \sum_{\tau=-\infty}^{\infty} |\text{cov}(X_t, X_\tau)| < \infty$, often called short memory), then the variance is $O\left(\frac{1}{n}\right)$, just as in the iid case. However, even with this assumption we need to be able to estimate $\text{var}(\bar{X})$ in order to test/construct CI for $\mu$. Usually this requires the stronger assumption of stationarity, which we define in Section 1.5.

Example 1.4.1 (The variance of a regression model with correlated errors) Let us return to the parametric models discussed in Section 1.2.1. The general model is

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

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 1.2.1. We use least squares to estimate $\beta$

$$\mathcal{L}_n(\beta) = \sum_{t=1}^n (Y_t - \beta' u_t)^2,$$
\begin{equation}
\hat{\beta}_n = \arg \min L_n(\beta) = \left( \sum_{t=1}^{n} u_t u_t' \right)^{-1} \sum_{t=1}^{n} Y_t u_t.
\end{equation}

Thus \( \frac{\partial L_n(\hat{\beta}_n)}{\partial \beta} = 0 \). To evaluate the variance of \( \hat{\beta}_n \) we will derive an expression for \( \hat{\beta}_n - \beta \) (this expression also applies to many nonlinear estimators too). We note that by using straightforward algebra we can show that

\begin{equation}
\frac{\partial L_n(\hat{\beta}_n)}{\partial \beta} - \frac{\partial L_n(\beta)}{\partial \beta} = \left[ \hat{\beta}_n - \beta \right]' \sum_{t=1}^{n} u_t u_t'. \tag{1.9}
\end{equation}

Moreover because \( \frac{\partial L_n(\hat{\beta}_n)}{\partial \beta} = 0 \) we have

\begin{align*}
\frac{\partial L_n(\hat{\beta}_n)}{\partial \beta} - \frac{\partial L_n(\beta)}{\partial \beta} &= - \frac{\partial L_n(\beta)}{\partial \beta} \\
&= \sum_{t=1}^{n} [Y_t - \beta' u_t] u_t = \sum_{t=1}^{n} u_t \varepsilon_t. \tag{1.10}
\end{align*}

Altogether (1.9) and (1.10) give

\begin{equation}
\left[ \hat{\beta}_n - \beta \right]' \sum_{t=1}^{n} u_t u_t' = \sum_{t=1}^{n} u_t' \varepsilon_t.
\end{equation}

and

\begin{equation}
\left[ \hat{\beta}_n - \beta \right] = \left( \sum_{t=1}^{n} u_t u_t' \right)^{-1} \sum_{t=1}^{n} u_t \varepsilon_t.
\end{equation}

Using this expression we can see that

\begin{equation}
\text{var} \left[ \hat{\beta}_n - \beta \right] = \left( \frac{1}{n} \sum_{t=1}^{n} u_t u_t' \right)^{-1} \text{var} \left( \frac{1}{n} \sum_{t=1}^{n} u_t \varepsilon_t \right) \left( \frac{1}{n} \sum_{t=1}^{n} u_t u_t' \right)^{-1}.
\end{equation}
Finally we need only evaluate \( \text{var} \left( \frac{1}{n} \sum_{t=1}^{n} u_t \varepsilon_t \right) \) which is

\[
\text{var} \left( \frac{1}{n} \sum_{t=1}^{n} u_t \varepsilon_t \right) = \frac{1}{n^2} \sum_{t,t'=1}^{n} \text{cov}[\varepsilon_t, \varepsilon_{t'}] u_t u'_{t'} \\
= \frac{1}{n^2} \sum_{t=1}^{n} \text{var}[\varepsilon_t] u_t u'_{t} + \frac{2}{n^2} \sum_{t=1}^{n} \sum_{\tau=t+1}^{n} \text{cov}[\varepsilon_t, \varepsilon_{\tau}] u_t u'_{\tau} .
\]

expression if independent additional term due to correlation in the errors

Under the assumption that \( \left( \frac{1}{n} \sum_{t=1}^{n} u_t u'_{t} \right) \) is non-singular, \( \sup_t \|u_t\|_1 < \infty \) and \( \sup_t \sum_{\tau=-\infty}^{\infty} |\text{cov}(\varepsilon_t, \varepsilon_{\tau})| < \infty \), we can see that \( \text{var} \left[ \hat{\beta}_n - \beta \right] = O(n^{-1}) \), but just as in the case of the sample mean we need to impose some additional conditions on \( \{\varepsilon_t\} \) if we want to construct confidence intervals/test \( \beta \).

### 1.5 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.
1.5.1 Types of stationarity (with Ergodicity thrown in)

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 1.5.1 (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 1.5.2 (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 1.5.1 (Strict and second order stationarity)**

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

\[
\text{cov}(X_t, X_{t+k}) = \mathbb{E}(X_t X_{t+k}) - \mathbb{E}(X_t)\mathbb{E}(X_{t+k})
\]

\[
= \int xy \left[ P_{X_t, X_{t+k}}(dx, dy) - P_{X_t}(dx)P_{X_{t+k}}(dy) \right]
\]

\[
= \int xy \left[ P_{X_0, X_k}(dx, dy) - P_{X_0}(dx)P_{X_k}(dy) \right] = \text{cov}(X_0, X_k),
\]

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 \( \text{cov}(X_t, X_{t+k}) \) 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 1.5.1 (The sample mean and its variance under stationarity) Returning the variance of the sample mean discussed (1.8), if a time series is second order stationary, then the sample mean $\bar{X}$ is estimating the mean $\mu$ and the variance of $\bar{X}$ is

$$\text{var}(\bar{X}) = \frac{1}{n} \text{var}(X_0) + \frac{2}{T^2} \sum_{t=1}^{n} \sum_{\tau=t+1}^{n} \text{cov}(X_t, X_\tau)$$

$$= \frac{1}{n} \text{var}(X_0) + \frac{2}{n} \sum_{r=1}^{n} \left( \frac{n-r}{n} \right) \text{cov}(X_0, X_r).$$

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 Chapter A) $\sum_{r=1}^{n} (1 - r/n)c(r) \to \sum_{r=1}^{\infty} c(r)$. This implies that

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

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

$$\text{E}(\bar{X} - \mu)^2 = \text{var}(\bar{X}) \to 0,$$

which we recall is convergence in mean square. Thus we have convergence in probability $\bar{X} \xrightarrow{P} \mu$.

The example above illustrates how second order stationarity gave a rather elegant expression for the variance. We now motivate the concept of ergodicity.

Sometimes, it is difficult to evaluate the mean and variance of an estimator, but often we may only require almost sure or convergence in probability. Therefore, we may want to find an alternative method to evaluating 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

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

35
and in general \( \frac{1}{n} \sum_{t=1}^{n} g(X_t) \overset{a.s.}{\to} \mathbb{E}[g(X_0)] \) (if \( \mathbb{E}[g(X_0)] < \infty \)). 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 1.5.3 (Ergodicity - very rough)** 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. Thus a process which is stationary is measure preserving.

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 \( \omega \) which generate sequences which ‘repeat’ itself (are periodic in some sense).

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

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

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

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

\[
\frac{1}{n} \sum_{t=1}^{n} g(X_t, X_{t+\tau_1}, \ldots, X_{t+\tau_k}) \overset{a.s.}{\to} \mathbb{E}[g(X_0, \ldots, X_{t+\tau_k})]
\] (1.11)
(often (1.11) is used as the definition of ergodicity, as it is an iff with the ergodic definition). Later you will see how useful this.

(1.11) 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

\[
X_t = h(..., \varepsilon_t, \varepsilon_{t-1}, ...),
\]

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 1.5.2** 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 \( E(X_t|Z = i) = \mu_i \) and \( E(X_t) = \frac{1}{2}(\mu_1 + \mu_2) \). 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 \( \frac{1}{2}(\mu_1 + \mu_2) \).

**Exercise 1.6** 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 distribution.

(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 \( E[U_t^2] < \infty \) and \( E[V_t^2] < \infty \).
Example 1.5.2 In Chapter 6 we consider estimation of the autocovariance function. However for now use the R command \texttt{acf}; it essentially evaluates \( \hat{\rho}(r) = \hat{c}(r)/\hat{c}(0) \), where

\[
\hat{c}(r) = \frac{1}{n} \sum_{t=1}^{n-r} (X_t - \bar{X})(X_{t+r} - \bar{X})
\]

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

In Figure 1.14 we give the sample acf plots of the Southern Oscillation Index and the Sunspot data. We observe that they 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 data appear to decay but over a longer period of time and there is a clear periodicity in the correlation.

![Figure 1.14: Top: ACF of Southern Oscillation data. Bottom ACF plot of Sunspot data.](image)

Exercise 1.7

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

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

(iii) Make an ACF plot of the residuals (after fitting a line through the data (using the command \texttt{lsfit(\ldots)$res})) of the yearly temperature data from 1880-2013.
R code

To make the above plots we use the commands

\[
\text{par(mfrow=c(2,1))}
\]
\[
\text{acf(soi,lag.max=300)}
\]
\[
\text{acf(sunspot,lag.max=60)}
\]

1.5.2 Towards statistical inference for time series

Returning to the sample mean Example 1.5.1. Suppose we want to construct CIs or apply statistical tests on the mean. This requires us to estimate the long run variance (assuming stationarity)

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

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.
1.6 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 1.6.1 (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 \( x = (x_1, \ldots, x_n) \in \mathbb{R}^n \) 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 \textit{even} 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 1.6.2 (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 \( x = (x_1, \ldots, x_n) \in \mathbb{R}^n \) 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 \textit{even} positive semi-definite function if (i) is satisfied and \( c(u) = c(-u) \) for all \( u \in \mathbb{R} \).

**Remark 1.6.1** You have probably encountered this positive definite notion before, when dealing with positive definite matrices. Recall the \( n \times n \) matrix \( \Sigma_n \) is positive semi-definite if for all \( x \in \mathbb{R}^n \)

\[
x' \Sigma_n x \geq 0.
\]

To see how this is related to positive semi-definite matrices, suppose that the matrix \( \Sigma_n \) has a special form, that is the elements of \( \Sigma_n \) are \( (\Sigma_n)_{i,j} = c(i - j) \). Then \( x' \Sigma_n x = \sum_{i,j} c(i - j)x_i x_j \).

We observe that in the case that \( \{X_t\} \) is a stationary process with covariance \( c(k) \), the variance covariance matrix of \( X_n = (X_1, \ldots, X_n) \) is \( \Sigma_n \), 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 1.6.1 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 even 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 \( x = (x_1, \ldots, x_n) \in \mathbb{R}^n \), 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 \( \text{var}(Y) = x' \text{var}(X_n)x = \sum_{i,j=1}^n c(i-j) x_i x_j \) where \( X_n = (X_1, \ldots, X_n) \). 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 straightforward, but depends on defining the characteristic function of a process and using Kolmogorov’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 1.6.1 there must exist a spatial process which has this covariance function). It is impossible to check for positive definiteness using Definitions 1.6.1 or 1.6.1. 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 1.6.2 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 and 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 \( \{c(u); u \in \mathbb{R}^d\} \) is a positive semi-definite sequence if and if

\[
 f(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} c(u) \exp(iu'\omega) du \geq 0
\]

for all \( \omega^d \in \mathbb{R}^d \).

**PROOF.** See Section 8.3.1.

**Example 1.6.1** 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 \( \varepsilon_t \) are iid random variables with \( \text{E}[\varepsilon_t] = 0 \) and \( \text{var}(\varepsilon_t) = 0.5 \). This model has this covariance structure.

We note that Theorem 1.6.2 can easily be generalized to higher dimensions, \( d \), by taking Fourier transforms over \( \mathbb{Z}^d \) or \( \mathbb{R}^d \).
Exercise 1.8 Which of these sequences can be 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 1.9

(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.