Chapter 1

Introduction

A time series is a series of observations $x_t$, each observed at the time $t$. 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. However in this course we will focus on the case that observations are observed at fixed time points, hence we will suppose we observe $\{x_t : t = 1, \ldots, n\}$. Below we give examples of typical time series. Figure 1.1 is of the daily exchange rate between the British pound and the US dollar (after taking log differences). Figure 1.2 is of the monthly minimum temperatures recorded at Antarctic and the Figure 1.3 is of the global temperature anomalies. Comparing the Antarctic, exchange rate data and global temperatures with the simulation of white noise (iid random variables) in Figure 1.4, we see that unlike the iid realisation, there appears to be ‘more smoothness’ in the plots and dependence between observations which are close located close in time. Figures 1.1, 1.2 and 1.3 are examples of time series and various time series models are fitted to this type of data.

Hence 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) $\Omega$ 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 $P(A)$ is known.

(iii) $P$ is the probability.

After all this formalisation, let us return to plots in Figures 1.2 and 1.1. We see that Figure 1.1 can be considered as one realisation from the stochastic process $\{X_t\}$. Now based on the one realisation
Figure 1.1: The GBP and USD exchange rate from 2000-2008 (after taking log differences)

Figure 1.2: The monthly minimum temperatures at Faraday station in the Antarctic.
Figure 1.3: The global yearly temperature anomalies from 1850-present

Figure 1.4: A simulation from 150 iid random variables
we want to make inference about parameters associated with the process \{X_t\}, such as the mean etc. Let us consider estimators of the mean, noting that the discussion below equally applies to any population parameter. 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 not independent and highly dependent, then it is likely that \{X_t\} would be concentrated about a small part of the probability space. In this case, the sample mean would be highly biased, and a consistent estimator is not possible.

A typical time series is a half way house between totally dependent data and independent data. For most time series parameter estimation is based on only one realisation \( x_t = X_t(\omega) \). 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 is stationary (this is a type of invariance assumption, that is the main character of the process do not change over time such as the mean does not change over time) (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 between in the stochastic process \{X_t\} is defined as

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

Hence if \{X_t\} has mean zero, then the above reduces to \( \text{cov}(X_t, X_{t+k}) = \mathbb{E}(X_t X_{t+k}) \). In a lot of statistical analysis the covariance is often sufficient as a measure. However it is worth bearing in mind that the covariance only measure linear dependence, usually 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. Examples of more general measures include mixing (in its various flavours), first introduced by Rosenblatt in the 50s (M. and Grenander (1997)). However in this course we will not cover mixing.

Returning to the sample mean example suppose that \{X_t\} is a time series with a constant mean \( \mu \), we estimate the mean \( \mu \) with the sample mean \( \bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t \). It is clear that this is an unbiased estimator of \( \mu \), since \( \mathbb{E}(\bar{X}) = \mu \), but to see whether it converges in probability to \( \mu \) we consider its variance

\[
\text{var}(\bar{X}) = \frac{1}{T^2} \sum_{t=1}^{T} \text{var}(X_t) + \frac{2}{T^2} \sum_{t>T} \text{cov}(X_t, X_{t'}). \]

If the covariance structure decays at such a rate that the sum of all lags is finite (\( \sup_{t} \sum_{t} |\text{cov}(X_t, X_{t'})| < \infty \), often called short memory), then the variance is \( O\left(\frac{1}{T}\right) \), 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 below.

**Data sets**

Websites where the data can be obtained include:

- [http://www.cru.uea.ac.uk/](http://www.cru.uea.ac.uk/)

### 1.1 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, and over shorter 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. 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 1.1.1 (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 data it is very hard to check for it). Often it is possible to work under a weaker assumption called weak stationarity.
Definition 1.1.2 (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 1.1.1 (i) It is easy to show that strict stationarity implies second order stationarity. But the converse is not necessarily true. To show that strict stationarity implies second order stationarity, suppose that \( \{X_t\} \) is a strictly stationary process with zero mean, then
\[
\text{cov}(X_t, X_{t+k}) = \int xy P_{X_t, X_{t+k}}(dx, dy) = \text{cov}(X_t, X_{t+k}),
\]
where \( P_{X_t, X_{t+k}} \) is the joint distribution of \( X_t, X_{t+k} \). Clearly \( \text{cov}(X_t, X_{t+k}) \) does not depend on \( t \) and \( \{X_t\} \) is second order stationary.

(ii) 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).

Returning the average discussed above, if a time series is second order stationary (and the covariances \( \sum_r |\text{cov}(X_0, X_r)| < \infty \), this for all \( r \), \( (1 - r/n)c(r) \to c(r) \) and \( \sum_r (1 - |r|/n)c(r) \leq \sum_r |c(r)| \), thus by dominated convergence \( \sum_{r=1}^{\infty} (1 - r/n)c(r) \to \sum_{r=1}^{\infty} c(r) \)), then the sample mean \( \bar{X} \) is estimating the mean \( \mu \) and the variance of \( \bar{X} \) is
\[
\text{var}(\bar{X}) = \frac{1}{T} \text{var}(X_0) + \frac{2}{T^2} \sum_{t=1}^{T} \sum_{\tau=t+1}^{T} \text{cov}(X_t, X_\tau)
\]
\[
= \frac{1}{T} \text{var}(X_0) + \frac{2}{T} \sum_{t=1}^{T} \left( \frac{T-t}{T} \right) \text{cov}(X_0, X_{\tau-t})
\]
\[
\approx \frac{1}{T} \text{var}(X_0) + \frac{2}{T} \sum_{t=1}^{\infty} \text{cov}(X_t, X_\tau) = O\left( \frac{1}{T} \right).
\]

The above is often called the long term variance. The above implies that
\[
\mathbb{E}(\bar{X} - \mu)^2 = \text{var}(\bar{X}) \to 0
\]
(convergence in mean square) and thus we have convergence in probability \( \bar{X} \xrightarrow{P} \mu \).

But sometimes it is a bit of a bother to evaluate the variance of an estimator, this raises the question of whether we can have almost sure (probability) convergence without 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
\[
\frac{1}{T} \sum_{t=1}^{T} X_t \xrightarrow{a.s.} \mu
\]
and in general \( \frac{1}{T} \sum_{t=1}^{T} g(X_t) \xrightarrow{a.s.} \mathbb{E}(g(X_0)) \). Does such a result exists in time series? It does, but we need to have the slightly stronger condition that a time series is ergodic (which is a slightly stronger condition than stationary).

**Definition 1.1.3 (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 clear 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 \((\Omega, \mathcal{F}, P)\).

The measure preserving transformation is

\[
T(x_1, x_2, \ldots) = (x_2, x_3, \ldots),
\]

strict stationarity basically allows it to have this measure preserving transformation. Thus the stochastic process \( \{X_t\} \) is said to be ergodic if a set \( A \in \mathcal{F} \) satisfies

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

then \( P(A) = 0 \) or \( 1 \). Roughly what 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 is quite complex and is rarely used in time series analysis (it is mainly used in probability theory and dynamical systems). However, one consequence of ergodicity is the ergodic theorem, which is extremely useful in time series. It states that \( \{X_t\} \) is an ergodic stochastic process then

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

for any function \( g(\cdot) \). Later you will see how useful this.

Here we see 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. That is \( \{Y_t\} \) where

\[
Y_t = g(\ldots, X_t, X_{t-1}, \ldots),
\]

so long as \( g \) is well defined (in the sense it has a finite limit), then \( Y_t \) is ergodic. For full details see Stout (1974).

**Remark 1.1.2** As mentioned above all ergodic processes are stationary, but a stationary process is not necessarily ergodic. The cases that a stationary process is not ergodic is rare but can arise. Here is one simple example. Suppose that \( \{\varepsilon_t\} \) are iid random variables and \( Z \) is a bernoulli random
variable with comes \(\{1, 2\}\) (where the chance of either outcome is half). Suppose that \(Z\) stays the same for all \(t\). Define
\[
X_t = \begin{cases} 
\mu_1 + \varepsilon_t & Z = 1 \\
\mu_2 + \varepsilon_t & Z = 2.
\end{cases}
\]
such that \(\mathbb{E}(X_t|Z = i) = \mu_i\) and \(\mathbb{E}(X_t) = \frac{1}{2}(\mu_1 + \mu_2)\). This sequence is stationary. In this example, we see \(\frac{1}{T}\sum_{t=1}^{T} X_t\) will only converge to one of the means, hence we do not have almost sure convergence to \(\frac{1}{2}(\mu_1 + \mu_2)\) (and also not convergence in probability).

However, often we need to construct CIs or do statistical tests. This requires us to estimate the variance (assuming stationarity)
\[
\text{var}(\bar{X}) \approx \frac{1}{T}\text{var}(X_0) + \frac{2}{T}\sum_{t=1}^{\infty}\text{cov}(X_t, X_r) = O\left(\frac{1}{T}\right).
\]
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.

Extras

The covariance of a stationary process has several very interesting properties. One of the main properties is that it is non-negative definite, which we define below.

**Definition 1.1.4** (Non-negative definite) A sequence \(\{c(k)\}\) is said to be non-negative 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
\[
\sum_{i,j}^n x_i c(i - j) x_j, \]
Remark 1.1.3 You have probably encountered this notion before when dealing with non-negative definite (positive definite) matrices. Recall the $n \times n$ matrix $\Sigma_n$ is non-negative definite if for all $x \in \mathbb{R}^n$ $x'\Sigma_n x \geq 0$. To see how this is related to non-negative 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 $x'\Sigma_n x = \sum_{i,j} x_i c(i-j) x_j$. 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 non-negative definite.

Theorem 1.1.1 Suppose that $\{X_t\}$ is a stationary time series with covariance function $\{c(k)\}$, then $\{c(k)\}$ is a non-negative definite sequence. Conversely for any non-negative definite sequence there exists a stationary time series with this non-negative definite sequence as its covariance function.

PROOF. To show that $\{c(k)\}$ is non-negative definite. Consider any sequence $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, and the double sum $\sum_{i,j} x_i c(i-j) x_j$. Define the random variable $Y = \sum_{i=1}^n x_i X_i$. It is straightforward to see that $\text{var}(Y) = x' \text{var}(X_n) x = \sum_{i,j} x_i c(i-j) x_j$ where $X_n = (X_1, \ldots, X_n)$. Since for any random variable $Y$, $\text{var}(Y) \geq 0$, this means that $\sum_{i,j} x_i c(i-j) x_j \geq 0$, hence $\{c(k)\}$ is a positive definite sequence.

To show the converse, that for any non-negative 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 Komologorov’s extension theorem. We omit the details but refer an interested reader to Brockwell and Davis (1998), Section 1.5. □

It is worth noting that a simple way to check for non-negative definiteness of sequence is to consider its Fourier transform. If the Fourier transform is positive, then the sequence is non-negative definite, we will look at this in more depth when we consider the spectral density.

The above theorem applies also to spatial processes. Which is why in spatial statistics they often look at the construction of positive definite covariance functions.