

Time Series Analysis

Lecture Notes for 475.726

Ross Ihaka

Statistics Department
University of Auckland

April 14, 2005

Contents

1	Introduction	1
1.1	Time Series	1
1.2	Stationarity and Non-Stationarity	1
1.3	Some Examples	2
1.3.1	Annual Auckland Rainfall	2
1.3.2	Nile River Flow	2
1.3.3	Yield on British Government Securities	2
1.3.4	Ground Displacement in an Earthquake	3
1.3.5	United States Housing Starts	3
1.3.6	Iowa City Bus Ridership	3
2	Vector Space Theory	7
2.1	Vectors In Two Dimensions	7
2.1.1	Scalar Multiplication and Addition	7
2.1.2	Norms and Inner Products	8
2.2	General Vector Spaces	9
2.2.1	Vector Spaces and Inner Products	9
2.2.2	Some Examples	11
2.3	Hilbert Spaces	12
2.3.1	Subspaces	13
2.3.2	Projections	13
2.4	Hilbert Spaces and Prediction	14
2.4.1	Linear Prediction	14
2.4.2	General Prediction	15
3	Time Series Theory	17
3.1	Time Series	17
3.2	Hilbert Spaces and Stationary Time Series	18
3.3	The Lag and Differencing Operators	19
3.4	Linear Processes	20
3.5	Autoregressive Series	21
3.5.1	The AR(1) Series	21
3.5.2	The AR(2) Series	23
3.5.3	Computations	26
3.6	Moving Average Series	29
3.6.1	The MA(1) Series	29
3.6.2	Invertibility	29
3.6.3	Computation	30

3.7	Autoregressive Moving Average Series	30
3.7.1	The ARMA(1,1) Series	31
3.7.2	The ARMA(p,q) Model	32
3.7.3	Computation	32
3.7.4	Common Factors	34
3.8	The Partial Autocorrelation Function	34
3.8.1	Computing the PACF	36
3.8.2	Computation	37
4	Identifying Time Series Models	39
4.1	ACF Estimation	39
4.2	PACF Estimation	42
4.3	System Identification	43
4.4	Model Generalisation	45
4.4.1	Non-Zero Means	45
4.4.2	Deterministic Trends	47
4.4.3	Models With Non-stationary AR Components	47
4.4.4	The Effect of Differencing	48
4.5	ARIMA Models	49
5	Fitting and Forecasting	51
5.1	Model Fitting	51
5.1.1	Computations	51
5.2	Assessing Quality of Fit	54
5.3	Residual Correlations	58
5.4	Forecasting	59
5.4.1	Computation	59
5.5	Seasonal Models	61
5.5.1	Purely Seasonal Models	61
5.5.2	Models with Short-Term and Seasonal Components	63
5.5.3	A More Complex Example	66
6	Frequency Domain Analysis	75
6.1	Some Background	75
6.1.1	Complex Exponentials, Sines and Cosines	75
6.1.2	Properties of Cosinusoids	76
6.1.3	Frequency and Angular Frequency	77
6.1.4	Invariance and Complex Exponentials	77
6.2	Filters and Filtering	78
6.2.1	Filters	78
6.2.2	Transfer Functions	79
6.2.3	Filtering Sines and Cosines	80
6.2.4	Filtering General Series	80
6.2.5	Computing Transfer Functions	81
6.2.6	Sequential Filtering	81
6.3	Spectral Theory	82
6.3.1	The Power Spectrum	82
6.3.2	The Cramér Representation	83
6.3.3	Using The Cramér Representation	85
6.3.4	Power Spectrum Examples	86

6.4	Statistical Inference	88
6.4.1	Some Distribution Theory	88
6.4.2	The Periodogram and its Distribution	90
6.4.3	An Example – Sunspot Numbers	91
6.4.4	Estimating The Power Spectrum	92
6.4.5	Tapering and Prewhitening	95
6.4.6	Cross Spectral Analysis	96
6.5	Computation	99
6.5.1	A Simple Spectral Analysis Package for R	99
6.5.2	Power Spectrum Estimation	99
6.5.3	Cross-Spectral Analysis	100
6.6	Examples	100

Chapter 1

Introduction

1.1 Time Series

Time series arise as recordings of processes which vary over time. A recording can either be a continuous trace or a set of discrete observations. We will concentrate on the case where observations are made at discrete equally spaced times. By appropriate choice of origin and scale we can take the observation times to be $1, 2, \dots, T$ and we can denote the observations by Y_1, Y_2, \dots, Y_T .

There are a number of things which are of interest in time series analysis. The most important of these are:

Smoothing: The observed Y_t are assumed to be the result of “noise” values ε_t additively contaminating a smooth signal η_t .

$$Y_t = \eta_t + \varepsilon_t$$

We may wish to recover the values of the underlying η_t .

Modelling: We may wish to develop a simple mathematical model which explains the observed pattern of Y_1, Y_2, \dots, Y_T . This model may depend on unknown parameters and these will need to be estimated.

Forecasting: On the basis of observations Y_1, Y_2, \dots, Y_T , we may wish to predict what the value of Y_{T+L} will be ($L \geq 1$), and possibly to give an indication of what the uncertainty is in the prediction.

Control: We may wish to intervene with the process which is producing the Y_t values in such a way that the future values are altered to produce a favourable outcome.

1.2 Stationarity and Non-Stationarity

A key idea in time series is that of *stationarity*. Roughly speaking, a time series is stationary if its behaviour does not change over time. This means, for example, that the values always tend to vary about the same level and that their variability is constant over time. Stationary series have a rich theory and

their behaviour is well understood. This means that they play a fundamental role in the study of time series.

Obviously, not all time series that we encounter are stationary. Indeed, non-stationary series tend to be the rule rather than the exception. However, many time series are related in simple ways to series which are stationary. Two important examples of this are:

Trend models : The series we observe is the sum of a deterministic *trend* series and a stationary *noise* series. A simple example is the linear trend model:

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

Another common trend model assumes that the series is the sum of a periodic “seasonal” effect and stationary *noise*. There are many other variations.

Integrated models : The time series we observe satisfies

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

where ε_t is a stationary series. A particularly important model of this kind is the *random walk*. In that case, the ε_t values are independent “shocks” which perturb the current state Y_t by an amount ε_{t+1} to produce a new state Y_{t+1} .

1.3 Some Examples

1.3.1 Annual Auckland Rainfall

Figure 1.1 shows the annual amount of rainfall in Auckland for the years from 1949 to 2000. The general pattern of rainfall looks similar throughout the record, so this series could be regarded as being stationary. (There is a hint that rainfall amounts are declining over time, but this type of effect can occur over shortish time spans for stationary series.)

1.3.2 Nile River Flow

Figure 1.2 shows the flow volume of the Nile at Aswan from 1871 to 1970. These are yearly values. The general pattern of this data does not change over time so it can be regarded as stationary (at least over this time period).

1.3.3 Yield on British Government Securities

Figure 1.3 shows the percentage yield on British Government securities, monthly over a 21 year period. There is a steady long-term increase in the yields. Over the period of observation a trend-plus-stationary series model looks like it might be appropriate. An integrated stationary series is another possibility.

1.3.4 Ground Displacement in an Earthquake

Figure 1.4 shows one component of the horizontal ground motion resulting from an earthquake. The initial motion (a little after 4 seconds) corresponds to the arrival of the p -wave and the large spike just before six seconds corresponds to the arrival of the s -wave. Later features correspond to the arrival of surface waves. This is an example of a transient signal and cannot have techniques appropriate for stationary series applied to it.

1.3.5 United States Housing Starts

Figure 1.5 shows the monthly number of housing starts in the United States (in thousands). Housing starts are a leading economic indicator. This means that an increase in the number of housing starts indicates that economic growth is likely to follow and a decline in housing starts indicates that a recession may be on the way.

1.3.6 Iowa City Bus Ridership

Figure 1.6 shows the monthly average weekday bus ridership for Iowa City over the period from September 1971 to December 1982. There is clearly a strong seasonal effect superimposed on top of a general upward trend.

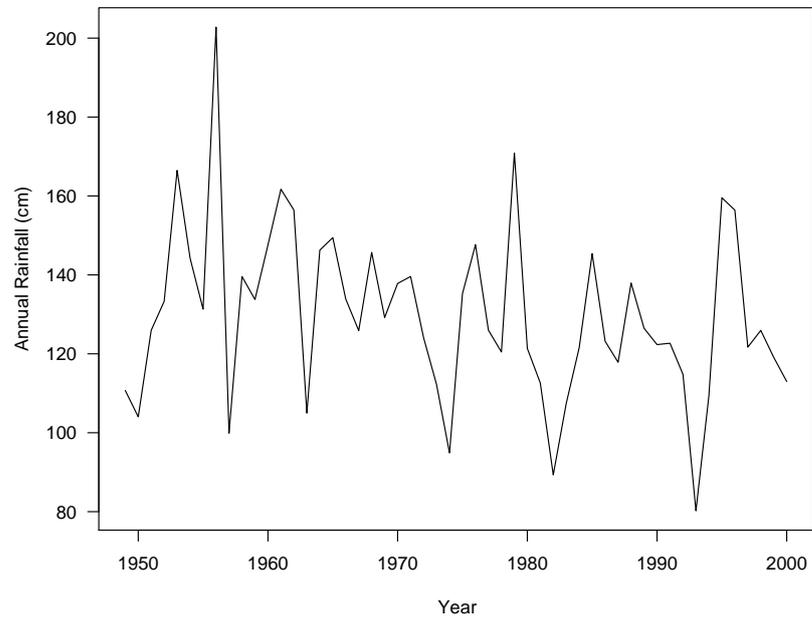


Figure 1.1: Annual Auckland rainfall (in cm) from 1949 to 2000 (from Paul Cowpertwait).

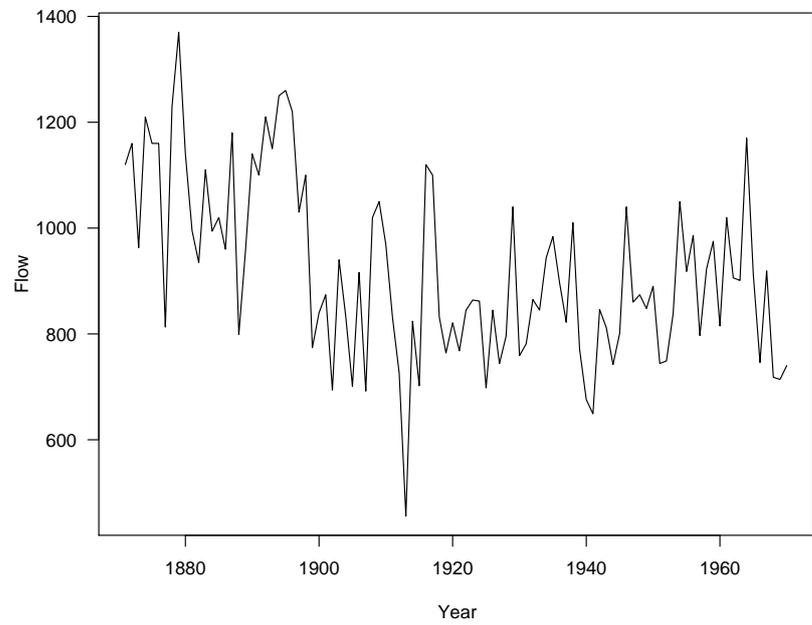


Figure 1.2: Flow volume of the Nile at Aswan from 1871 to 1970 (from Durbin and Koopman).

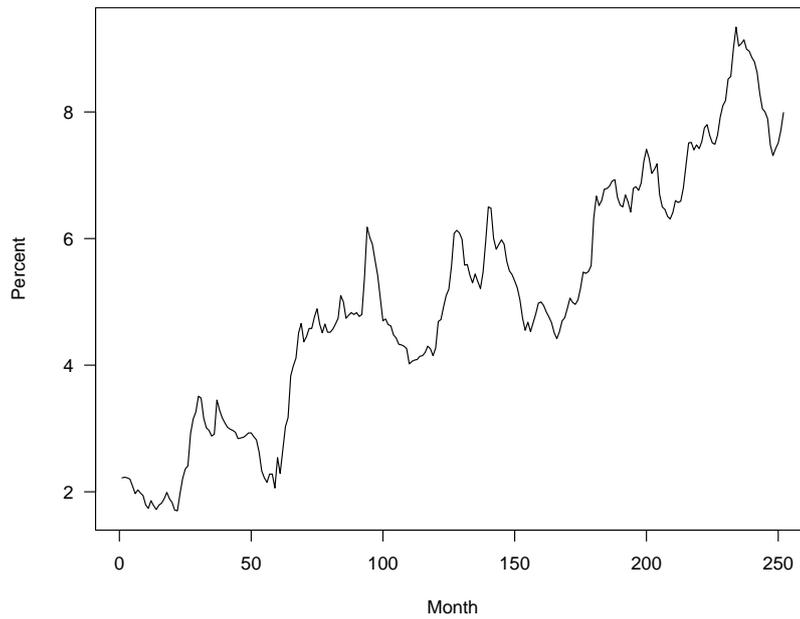


Figure 1.3: Monthly percentage yield on British Government securities over a 21 year period (from Chatfield).

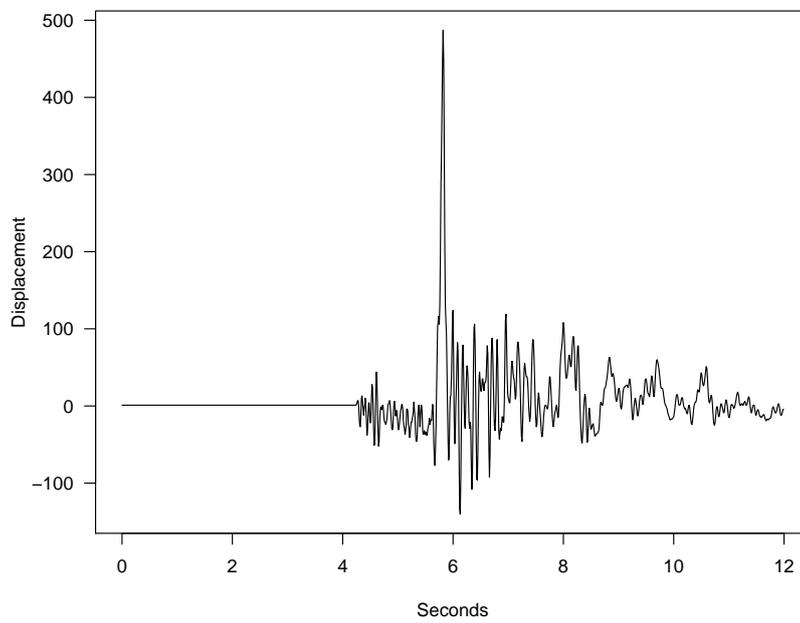


Figure 1.4: Horizontal ground displacement during a small Nevada earthquake (from Bill Peppin).

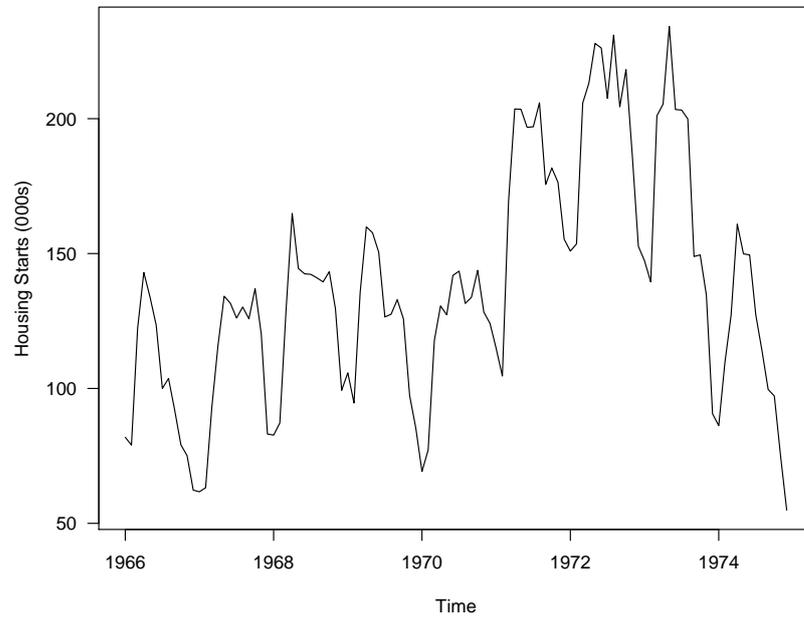


Figure 1.5: Housing starts in the United States (000s) (from S-Plus).

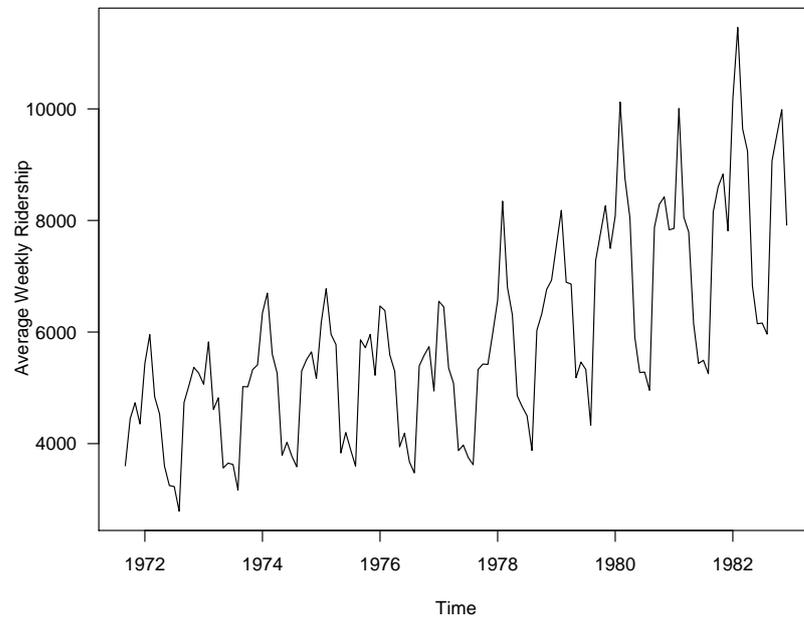


Figure 1.6: Average weekday bus ridership, Iowa City (monthly ave) Sep 1971 - Dec 1982.

Chapter 2

Vector Space Theory

2.1 Vectors In Two Dimensions

The theory which underlies time series analysis is quite technical in nature. In spite of this, a good deal of intuition can be developed by approaching the subject geometrically. The geometric approach is based on the ideas of *vectors* and *vector spaces*.

2.1.1 Scalar Multiplication and Addition

A good deal can be learnt about the theory of vectors by considering the two-dimensional case. You can think of two dimensional vectors as being little arrows which have a length and a direction. Individual vectors can be stretched (altering their length but not their direction) and pairs of vectors can be added by placing them head to tail.

To get more precise, we'll suppose that a vector \mathbf{v} extends for a distance x in the horizontal direction and a distance y in the vertical direction. This gives us a representation of the vector as a pair of numbers and we can write it as

$$\mathbf{v} = (x, y).$$

Doubling the length of the vector doubles the x and y values. In a similar way we can scale the length of the vector by any value, and this results in a similar scaling of the x and y values. This gives us a natural way of defining multiplication of a vector by a number.

$$c\mathbf{v} = (cx, cy)$$

A negative value for c produces a reversal of direction as well as change of length of magnitude $|c|$.

Adding two vectors amounts to placing them head to tail and taking the sum to be the arrow which extends from the base of the first to the head of the second. This corresponds to adding the x and y values corresponding to the vectors. For two vectors $\mathbf{v}_1 = (x_1, y_1)$ and $\mathbf{v}_2 = (x_2, y_2)$ the result is $(x_1 + x_2, y_1 + y_2)$. This corresponds to a natural definition of addition for vectors.

$$\mathbf{v}_1 + \mathbf{v}_2 = (x_1 + x_2, y_1 + y_2)$$

2.1.2 Norms and Inner Products

While coordinates give a complete description of vectors, it is often more useful to describe them in terms of the lengths of individual vectors and the angles between pairs of vectors. If we denote the length of a vector by $\|\mathbf{u}\|$, then by Pythagoras' theorem we know

$$\|\mathbf{u}\| = \sqrt{x^2 + y^2}.$$

Vector length has a number of simple properties:

Positivity: For every vector \mathbf{u} , we must have $\|\mathbf{u}\| \geq 0$, with equality if and only if $\mathbf{u} = \mathbf{0}$.

Linearity: For every scalar c and vector \mathbf{u} we have $\|c\mathbf{u}\| = |c|\|\mathbf{u}\|$.

The Triangle Inequality: If \mathbf{u} and \mathbf{v} are vectors, then $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.

The first and second of these properties are obvious, and the third is simply a statement that the shortest distance between two points is a straight line. The technical mathematical name for the length of a vector is the *norm* of the vector.

Although the length of an individual vector gives some information about it, it is also important to consider the angles between pairs of vectors. Suppose that we have vectors $\mathbf{v}_1 = (x_1, y_1)$ and $\mathbf{v}_2 = (x_2, y_2)$, and that we want to know the angle between them. If \mathbf{v}_1 subtends an angle θ_1 with the x axis and \mathbf{v}_2 subtends an angle θ_2 with the x axis, simple geometry tells us that:

$$\begin{aligned}\cos \theta_i &= \frac{x_i}{\|\mathbf{v}_i\|}, \\ \sin \theta_i &= \frac{y_i}{\|\mathbf{v}_i\|}.\end{aligned}$$

The angle we are interested in is $\theta_1 - \theta_2$ and the cosine of this can be computed using the formulae:

$$\begin{aligned}\cos(\alpha - \beta) &= \cos \alpha \cos \beta + \sin \alpha \sin \beta \\ \sin(\alpha - \beta) &= \sin \alpha \cos \beta - \cos \alpha \sin \beta,\end{aligned}$$

so that

$$\begin{aligned}\cos(\theta_1 - \theta_2) &= \frac{x_1x_2 + y_1y_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|} \\ \sin(\theta_1 - \theta_2) &= \frac{x_2y_1 - x_1y_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}.\end{aligned}$$

Knowledge of the sine and cosine values makes it possible to compute the angle between the vectors.

There are actually two angles between a pair of vectors; one measured clockwise and one measured counter-clockwise. Often we are interested in the smaller of these two angles. This can be determined from the cosine value (because $\cos \theta = \cos(-\theta)$). The cosine is determined by the lengths of the two vectors and the quantity $x_1x_2 + y_1y_2$. This quantity is called the *inner product* of the vectors \mathbf{v}_1 and \mathbf{v}_2 and is denoted by $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$. Inner products have the following basic properties.

Positivity: For every vector \mathbf{v} , we must have $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$, with $\langle \mathbf{v}, \mathbf{v} \rangle = 0$ only when $\mathbf{v} = \mathbf{0}$.

Linearity: For all vectors \mathbf{u} , \mathbf{v} , and \mathbf{w} ; and scalars α and β ; we must have $\langle \alpha\mathbf{u} + \beta\mathbf{v}, \mathbf{w} \rangle = \alpha\langle \mathbf{u}, \mathbf{w} \rangle + \beta\langle \mathbf{v}, \mathbf{w} \rangle$.

Symmetry: For all vectors \mathbf{u} and \mathbf{v} , we must have $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$ (or in the case of complex vector spaces $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$, where the bar indicates a complex conjugate).

It is clear that the norm can be defined in terms of the inner product by

$$\|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle.$$

It is easy to show that the properties of the norm $\|\mathbf{v}\|$ follow from those of the inner product.

Two vectors \mathbf{u} and \mathbf{v} are said to be *orthogonal* if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. This means that the vectors are “at right angles” to each other.

2.2 General Vector Spaces

The theory and intuition obtained from studying two-dimensional vectors carries over into more general situations. It is not the particular representation of vectors as pairs of coordinates which is important but rather the ideas of addition, scalar multiplication and inner-product which are important.

2.2.1 Vector Spaces and Inner Products

The following concepts provide an abstract generalisation of vectors in two dimensions.

1. A vector space is a set of objects, called vectors, which is closed under addition and scalar multiplication. This means that when vectors are scaled or added, the result is a vector.
2. An inner product on a vector space is a function which takes two vectors \mathbf{u} and \mathbf{v} and returns a scalar denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$ which has the conditions of positivity, linearity and symmetry described in section 2.1.2. A vector space with an associated inner product is called an inner product space.
3. The norm associated with an inner product space is defined in terms of the inner product as $\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$.

These ideas can be applied to quite general vector spaces. Although it may seem strange to apply ideas of direction and length to some of these spaces, thinking in terms of two and three dimensional pictures can be quite useful.

Example 2.2.1 The concept of two-dimensional vectors can be generalised by looking at the set of n -tuples of the form $\mathbf{u} = (u_1, u_2, \dots, u_n)$, where each u_i is a real number. With addition and scalar multiplication defined element-wise,

the set of all such n -tuples forms a vector space, usually denoted by \mathbb{R}^n . An inner product can be defined on the space by

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^n u_i v_i,$$

and this produces the norm

$$\|\mathbf{u}\| = \left(\sum_{i=1}^n u_i^2 \right)^{\frac{1}{2}}.$$

This generalisation of vector ideas to n dimensions provides the basis for a good deal of statistical theory. This is especially true for linear models and multivariate analysis.

We will need a number of fundamental results on vector spaces. These follow directly from the definition of vector space, inner-product and norm, and do not rely on any special assumptions about the kind of vectors being considered.

The Cauchy-Schwarz Inequality. For any two vectors \mathbf{u} and \mathbf{v}

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|. \quad (2.1)$$

Proof: For any two vectors \mathbf{u} and \mathbf{v} and scalars α and β ,

$$0 \leq \|\alpha \mathbf{u} - \beta \mathbf{v}\|^2 = \langle \alpha \mathbf{u} - \beta \mathbf{v}, \alpha \mathbf{u} - \beta \mathbf{v} \rangle = \alpha^2 \|\mathbf{u}\|^2 - 2\alpha\beta \langle \mathbf{u}, \mathbf{v} \rangle + \beta^2 \|\mathbf{v}\|^2$$

Setting $\alpha = \|\mathbf{v}\|$ and $\beta = \langle \mathbf{u}, \mathbf{v} \rangle / \|\mathbf{v}\|$ yields

$$0 \leq \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - 2\langle \mathbf{u}, \mathbf{v} \rangle^2 + \langle \mathbf{u}, \mathbf{v} \rangle^2 = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2.$$

This can be rewritten as

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|.$$

The Triangle Inequality. For any two vectors \mathbf{u} and \mathbf{v} ,

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$$

Proof: For any \mathbf{u} and \mathbf{v}

$$\begin{aligned} \|\mathbf{u} + \mathbf{v}\|^2 &= \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \\ &\leq \|\mathbf{u}\|^2 + 2\|\mathbf{u}\| \|\mathbf{v}\| + \|\mathbf{v}\|^2 \quad (\text{Cauchy-Schwarz}) \\ &= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2 \end{aligned}$$

The Angle Between Vectors. The Cauchy-Schwarz inequality means that

$$-1 \leq \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|} \leq 1.$$

This means that just as in the 2- d case we can define the angle between vectors \mathbf{u} and \mathbf{v} by

$$\theta = \cos^{-1} \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

2.2.2 Some Examples

The ideas in abstract vector space theory are derived from the concrete ideas of vectors in 2 and 3 dimensions. The real power of vector space theory is that it applies to much more general spaces. To show off the generality of the theory we'll now look at a couple of examples of less concrete spaces.

Example 2.2.2 The set \mathcal{F} of all (real-valued) random variables with $EX = 0$ and $E|X|^2 < \infty$ is a vector space, because for any choice of X_1 and X_2 from \mathcal{F} and scalars β_1 and β_2

$$E[\beta_1 X_1 + \beta_2 X_2] = 0$$

and

$$E|\beta_1 X_1 + \beta_2 X_2|^2 < \infty.$$

It is possible to define an inner product on this vector space by

$$\langle X, Y \rangle = EXY.$$

This in turn produces the norm

$$\|X\| = (E|X|^2)^{\frac{1}{2}},$$

which is just the standard deviation of X .

The cosine of the angle between random variables is defined as

$$\frac{EXY}{\sqrt{E|X|^2 E|Y|^2}},$$

which is recognisable as the correlation between the X and Y .

(There is a technical problem with uniqueness in this case. Any random variable which has probability 1 of being zero will have $\langle X, X \rangle = 0$, which violates the requirement that this only happen for $X = 0$. The workaround is to regard random variables as identical if they are the same with probability one.)

Example 2.2.3 The set of all continuous functions from $[-1, 1]$ to \mathbb{R} is a vector space because there is a natural definition of addition and scalar multiplication for such functions.

$$\begin{aligned}(f + g)(x) &= f(x) + g(x) \\ (\beta f)(x) &= \beta f(x)\end{aligned}$$

A natural inner-product can be defined on this space by

$$\langle f, g \rangle = \int_{-1}^1 f(x)g(x) dx.$$

Vector space theory immediately gets us results such as

$$\left| \int_{-1}^1 f(x)g(x) dx \right| \leq \left(\int_{-1}^1 f(x)^2 dx \right)^{1/2} \left(\int_{-1}^1 g(x)^2 dx \right)^{1/2},$$

which is just a restatement of the Cauchy-Schwarz inequality.

There is also a technical uniqueness problem in this case. This is handled by regarding functions f and g as equal if

$$\int_{-1}^1 |f(x) - g(x)|^2 dx = 0.$$

This happens, for example, if f and g differ only on a finite or countably infinite set of points. Integration theory gives a characterisation in terms of sets with “measure zero.”

2.3 Hilbert Spaces

The vector spaces we’ve looked at so far work perfectly well for performing finite operations such as forming linear combinations. However, there can be problems when considering limits.

Consider the case of example 2.2.3. Each of the functions f_n defined by

$$f_n(x) = \begin{cases} 0, & x < -\frac{1}{n}, \\ \frac{nx+1}{2}, & -\frac{1}{n} \leq x \leq \frac{1}{n}, \\ 1, & x > \frac{1}{n}, \end{cases}$$

is continuous and the sequence obviously converges to the function

$$f(x) = \begin{cases} 0, & x < 0, \\ 1/2, & x = 0, \\ 1, & x > 0. \end{cases}$$

This limit function is not in the space under consideration because it is not continuous.

Hilbert spaces add a requirement of completeness to those for an inner product space. In Hilbert spaces, sequences that look like they are converging will actually converge to an element of the space. To make this precise we need to define the meaning of convergence.

Convergence. Suppose that $\{\mathbf{u}_n\}$ is a sequence of vectors and \mathbf{u} is a vector such that $\|\mathbf{u}_n - \mathbf{u}\| \rightarrow 0$, then \mathbf{u}_n is said to converge to \mathbf{u} and this is denoted by $\mathbf{u}_n \rightarrow \mathbf{u}$.

If $\{\mathbf{u}_n\}$ is a sequence of vectors such that the $\mathbf{u}_n \rightarrow \mathbf{u}$ and \mathbf{v} is any vector then, by the Cauchy-Schwarz inequality $\langle \mathbf{u}_n, \mathbf{v} \rangle \rightarrow \langle \mathbf{u}, \mathbf{v} \rangle$. In a similar way $\|\mathbf{u}_n\| \rightarrow \|\mathbf{u}\|$. These properties are referred to as continuity of the inner product and norm.

A sequence for which $\lim_{m,n \rightarrow \infty} \|\mathbf{u}_m - \mathbf{u}_n\| \rightarrow 0$ is called a *Cauchy sequence*. It is easy to see that every convergent sequence is a Cauchy sequence. If conversely every Cauchy sequence converges to an element of the space, the space is said to be *complete*. A complete inner product space is called a *Hilbert space*.

Hilbert spaces preserve many of the important properties of \mathbb{R}^n . In particular the notions of length and direction retain their intuitive meanings. This makes it possible to carry out mathematical arguments geometrically and even to use pictures to understand what is happening in quite complex cases.

2.3.1 Subspaces

A subset \mathcal{M} of a Hilbert space \mathcal{H} is called a *linear manifold* if whenever \mathbf{u} and \mathbf{v} are elements of \mathcal{M} then so is $\alpha\mathbf{u} + \beta\mathbf{v}$. A linear manifold which contains the limit of every Cauchy sequence of its elements is called a linear *subspace* of \mathcal{H} . A linear subspace of a Hilbert space is itself a Hilbert space.

A vector \mathbf{v} is orthogonal to a subspace \mathcal{M} if $\langle \mathbf{v}, \mathbf{u} \rangle = 0$ for every $\mathbf{u} \in \mathcal{M}$. The set all vectors which are orthogonal to \mathcal{M} is itself a subspace denoted by \mathcal{M}^\perp and called the *orthogonal complement* of \mathcal{M} .

A set of vectors $\{\mathbf{v}_\lambda : \lambda \in \Lambda\}$ is said to generate a linear subspace \mathcal{M} if \mathcal{M} is the smallest subspace containing those vectors. It is relatively straightforward to show that the subspace consists of all linear combinations of the \mathbf{v}_λ s together with the limits of all Cauchy sequences formed from these linear combinations.

We will use the notation $\overline{\text{sp}}\{\mathbf{v}_\lambda : \lambda \in \Lambda\}$ to indicate the subspace generated by the random variables $\{\mathbf{v}_\lambda : \lambda \in \Lambda\}$.

2.3.2 Projections

If \mathcal{M} is a subspace of a Hilbert space \mathcal{H} and \mathbf{v} is a vector not in \mathcal{M} then the distance from \mathcal{M} to \mathbf{v} is defined to be $\min_{\mathbf{u} \in \mathcal{M}} \|\mathbf{v} - \mathbf{u}\|$. A crucial result in Hilbert space theory tells us about how this minimum is attained.

The Projection Theorem. There is a unique vector $\mathcal{P}_\mathcal{M}\mathbf{v} \in \mathcal{M}$ such that

$$\|\mathbf{v} - \mathcal{P}_\mathcal{M}\mathbf{v}\| = \min_{\mathbf{u} \in \mathcal{M}} \|\mathbf{v} - \mathbf{u}\|.$$

The vector $\mathcal{P}_\mathcal{M}\mathbf{v}$ satisfies $\mathcal{P}_\mathcal{M}\mathbf{v} \in \mathcal{M}$ and $\mathbf{v} - \mathcal{P}_\mathcal{M}\mathbf{v} \in \mathcal{M}^\perp$. It is called the *orthogonal projection* of \mathbf{v} on \mathcal{M} .

When \mathcal{M} is generated by a set of elements $\{\mathbf{u}_\lambda : \lambda \in \Lambda\}$, the condition $\mathbf{v} - \mathcal{P}_\mathcal{M}\mathbf{v} \in \mathcal{M}^\perp$ is equivalent to the condition $(\mathbf{v} - \mathcal{P}_\mathcal{M}\mathbf{v}) \perp \mathbf{u}_\lambda$ for every $\lambda \in \Lambda$. In other words, $\langle \mathbf{v} - \mathcal{P}_\mathcal{M}\mathbf{v}, \mathbf{u}_\lambda \rangle = 0$ for every $\lambda \in \Lambda$. This produces the equations

$$\langle \mathcal{P}_\mathcal{M}\mathbf{v}, \mathbf{u}_\lambda \rangle = \langle \mathbf{v}, \mathbf{u}_\lambda \rangle, \quad \lambda \in \Lambda,$$

which together with the requirement $\mathcal{P}_\mathcal{M}\mathbf{v} \in \mathcal{M}$, completely determines $\mathcal{P}_\mathcal{M}\mathbf{v}$.

When a subspace of a Hilbert space is generated by a countable set of orthogonal vectors $\{\boldsymbol{\xi}_n\}$, the orthogonal projection has the form

$$\mathcal{P}_\mathcal{M}\mathbf{v} = \sum_n \frac{\langle \mathbf{v}, \boldsymbol{\xi}_n \rangle}{\|\boldsymbol{\xi}_n\|^2} \boldsymbol{\xi}_n$$

In the case of a finite set $\{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n\}$ this is easy to see.

$$\frac{\langle \mathbf{v}, \boldsymbol{\xi}_1 \rangle}{\|\boldsymbol{\xi}_1\|^2} \boldsymbol{\xi}_1 + \dots + \frac{\langle \mathbf{v}, \boldsymbol{\xi}_n \rangle}{\|\boldsymbol{\xi}_n\|^2} \boldsymbol{\xi}_n$$

In addition, the orthogonality of $\{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n\}$ means

$$\begin{aligned} \left\langle \frac{\langle \mathbf{v}, \boldsymbol{\xi}_1 \rangle}{\|\boldsymbol{\xi}_1\|^2} \boldsymbol{\xi}_1 + \dots + \frac{\langle \mathbf{v}, \boldsymbol{\xi}_n \rangle}{\|\boldsymbol{\xi}_n\|^2} \boldsymbol{\xi}_n, \boldsymbol{\xi}_i \right\rangle &= \frac{\langle \mathbf{v}, \boldsymbol{\xi}_i \rangle}{\|\boldsymbol{\xi}_i\|^2} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_i \rangle \\ &= \langle \mathbf{v}, \boldsymbol{\xi}_i \rangle \end{aligned}$$

so the conditions above are verified.

The general (countable) case follows from the continuity of the norm and inner product.

2.4 Hilbert Spaces and Prediction

Consider the Hilbert space \mathcal{H} consisting of all finite-variance random variables on some probability space, with inner product defined by

$$\langle X, Y \rangle = \mathbf{E}XY.$$

We will now look at the problem of predicting a variable Y using zero, one or more predicting random variables X_1, \dots, X_n .

2.4.1 Linear Prediction

The problem can be stated as follows. Given a “response” random variable Y and predictor random variables X_1, \dots, X_n , what is the best way of predicting Y using a linear function of the X s. This amounts to finding the coefficients which minimise the mean squared error

$$\mathbf{E}|Y - \beta_0 - \beta_1 X_1 - \dots - \beta_n X_n|^2,$$

or, in a Hilbert space setting,

$$\|Y - \beta_0 - \beta_1 X_1 - \dots - \beta_n X_n\|^2.$$

The variables $\{1, X_1, \dots, X_n\}$ generate a subspace \mathcal{M} of \mathcal{H} , and the minimisation problem above amounts to finding the projection $\mathcal{P}_{\mathcal{M}}Y$. We’ll approach this problem in steps, beginning with $n = 0$.

The subspace \mathcal{C} generated by the constant random variable 1 consists of all constant random variables. Using the result above, the projection of a random variable Y with mean μ_Y and variance σ_Y^2 onto this subspace is

$$\mathcal{P}_{\mathcal{C}}Y = \frac{\langle Y, 1 \rangle}{\|1\|^2} 1 = \mathbf{E}Y = \mu_Y.$$

This tells us immediately that the value of c which minimises $\mathbf{E}[Y - c]^2$ is μ_Y .

Now consider the subspace \mathcal{L} generated by 1 and a single random variable X with mean μ_X and variance σ_X^2 . This is clearly the same as the subspace generated by 1 and $X - \mathbf{E}X$. Since 1 and $X - \mathbf{E}X$ are orthogonal, we can use the projection result of the previous section to compute the projection of Y onto \mathcal{L} .

$$\begin{aligned} \mathcal{P}_{\mathcal{L}}Y &= \frac{\langle Y, 1 \rangle}{\|1\|^2} 1 + \frac{\langle Y, X - \mathbf{E}X \rangle}{\|X - \mathbf{E}X\|^2} (X - \mathbf{E}X) \\ &= \mathbf{E}Y + \frac{\langle Y - \mathbf{E}Y, X - \mathbf{E}X \rangle}{\|X - \mathbf{E}X\|^2} (X - \mathbf{E}X) + \frac{\langle \mathbf{E}Y, X - \mathbf{E}X \rangle}{\|X - \mathbf{E}X\|^2} (X - \mathbf{E}X) \\ &= \mathbf{E}Y + \frac{\langle Y - \mathbf{E}Y, X - \mathbf{E}X \rangle}{\|X - \mathbf{E}X\|^2} (X - \mathbf{E}X) \end{aligned}$$

because $\langle \mathbb{E}Y, X - \mathbb{E}X \rangle = 0$.

Now $\langle Y - \mathbb{E}Y, X - \mathbb{E}X \rangle = \text{cov}(Y, X)$ which is in turn equal to $\rho_{YX}\sigma_Y\sigma_X$, where ρ_{YX} is the correlation between Y and X . This means that

$$\mathcal{P}_{\mathcal{L}}Y = \mu_Y + \rho_{YX} \frac{\sigma_Y}{\sigma_X} (X - \mu_X).$$

$\mathcal{P}_{\mathcal{L}}Y$ is the best possible linear prediction of Y based on X , because among all predictions of the form $\beta_0 + \beta_1 X$, it is the one which minimises

$$\mathbb{E}(Y - \beta_0 - \beta_1 X)^2.$$

The general case of n predictors proceeds in exactly the same way, but is more complicated because we must use progressive orthogonalisation of the set of variables $\{1, X_1, \dots, X_n\}$. The final result is that the best predictor of Y is

$$\mathcal{P}_{\mathcal{L}}Y = \mu_Y + \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} (\mathbf{X} - \boldsymbol{\mu}_X),$$

where \mathbf{X} represents the variables X_1, \dots, X_n assembled into a vector, $\boldsymbol{\mu}_X$ is the vector made up of the means of the X s, $\boldsymbol{\Sigma}_{YX}$ is the vector of covariances between Y and each of the X s, and $\boldsymbol{\Sigma}_{XX}$ is the variance-covariance matrix of the X s.

2.4.2 General Prediction

It is clear that linear prediction theory can be developed using Hilbert space theory. What is a little less clear is that Hilbert space theory also yields a general non-linear prediction theory.

Linear prediction theory uses only linear information about the predictors and there is much more information available. In the one variable case, the additional information can be obtained by considering all possible (Borel) functions $\phi(X)$. These are still just random variables and so generate a subspace \mathcal{M} of \mathcal{H} . The projection onto this subspace gives the best possible predictor of Y based on X_1, \dots, X_n .

With some technical mathematics it is possible to show that this projection is in fact just the conditional expectation $\mathbb{E}[Y|X]$. More generally, it is possible to show that the best predictor of Y based on X_1, \dots, X_n is $\mathbb{E}[Y|X_1, \dots, X_n]$.

Although this is theoretically simple, it requires very strong assumptions about the distribution of Y and X_1, \dots, X_n to actually compute an explicit value for the prediction. The simplest assumption to make is that Y and X_1, \dots, X_n have a joint normal distribution. In this case, the general predictor and the linear one are identical.

Because of this it is almost always the case that linear prediction is preferred to general prediction.

Chapter 3

Time Series Theory

3.1 Time Series

We will assume that the time series values we observe are the realisations of random variables Y_1, \dots, Y_T , which are in turn part of a larger *stochastic process* $\{Y_t : t \in \mathbb{Z}\}$. It is this underlying process that will be the focus for our theoretical development.

Although it is best to distinguish the observed time series from the underlying stochastic process, the distinction is usually blurred and the term time series is used to refer to both the observations and the underlying process which generates them.

The mean and the variance of random variables have a special place in the theory of statistics. In time series analysis, the analogs of these are the *mean function* and the *autocovariance function*.

Definition 3.1.1 (Mean and Autocovariance Functions): The *mean function* of a time series is defined to be $\mu(t) = EY_t$ and the *autocovariance function* is defined to be $\gamma(s, t) = \text{cov}(Y_s, Y_t)$.

The mean and the autocovariance functions are fundamental parameters and it would be useful to obtain sample estimates of them. For general time series there are $2T + T(T - 1)/2$ parameters associated with Y_1, \dots, Y_T and it is not possible to estimate all these parameters from T data values.

To make any progress at all we must impose constraints on the time series we are investigating. The most common constraint is that of stationarity. There are two common definitions of stationarity.

Definition 3.1.2 (Strict Stationarity): A time series $\{Y_t : t \in \mathbb{Z}\}$ is said to be *strictly stationary* if for any $k > 0$ and any $t_1, \dots, t_k \in \mathbb{Z}$, the distribution of

$$(Y_{t_1}, \dots, Y_{t_k})$$

is the same as that for

$$(Y_{t_1+u}, \dots, Y_{t_k+u})$$

for every value of u .

This definition says that the stochastic behaviour of the process does not change through time. If Y_t is stationary then

$$\mu(t) = \mu(0)$$

and

$$\gamma(s, t) = \gamma(s - t, 0).$$

So for stationary series, the mean function is constant and the autocovariance function depends only on the time-lag between the two values for which the covariance is being computed.

These two restrictions on the mean and covariance functions are enough for a reasonable amount of theory to be developed. Because of this a less restrictive definition of stationarity is often used in place of strict stationarity.

Definition 3.1.3 (Weak Stationarity): A time series is said to be *weakly, wide-sense* or *covariance stationary* if $E|Y_t|^2 < \infty$, $\mu(t) = \mu$ and $\gamma(t + u, t) = \gamma(u, 0)$ for all t and u .

In the case of Gaussian time series, the two definitions of stationarity are equivalent. This is because the finite dimensional distributions of the time series are completely characterised by the mean and covariance functions.

When time series are stationary it is possible to simplify the parameterisation of the mean and autocovariance functions. In this case we can define the mean of the series to be $\mu = E(Y_t)$ and the autocovariance function to be $\gamma(u) = \text{cov}(Y_{t+u}, Y_t)$. We will also have occasion to examine the *autocorrelation function*

$$\rho(u) = \frac{\gamma(u)}{\gamma(0)} = \text{cor}(Y_{t+u}, Y_t).$$

Example 3.1.1 (White Noise) If the random variables which make up $\{Y_t\}$ are uncorrelated, have means 0 and variance σ^2 , then $\{Y_t\}$ is stationary with autocovariance function

$$\gamma(u) = \begin{cases} \sigma^2 & u = 0, \\ 0 & \text{otherwise.} \end{cases}$$

This type of series is referred to as *white noise*.

3.2 Hilbert Spaces and Stationary Time Series

Suppose that $\{Y_t : t \in \mathbb{Z}\}$ is a stationary zero-mean time series. We can consider the Hilbert space \mathcal{H} generated by the random variables $\{Y_t : t \in \mathbb{Z}\}$ with inner product

$$\langle X, Y \rangle = E(XY),$$

and norm

$$\|X\|^2 = E|X|^2.$$

At a given time t , we can consider the subspace \mathcal{M} generated by the random variables $\{Y_u : u \leq t\}$. This subspace represents the past and present of the process. Future values of the series can be predicted by projecting onto the

subspace \mathcal{M} . For example, Y_{t+1} can be predicted by $\mathcal{P}_{\mathcal{M}}Y_{t+1}$, Y_{t+1} by $\mathcal{P}_{\mathcal{M}}Y_{t+2}$ and so on.

Computing these predictions requires a knowledge of the autocovariance function of the time series and typically this is not known. We will spend a good deal of this chapter studying simple parametric models for time series. By fitting such models we will be able to determine the covariance structure of the time series and so be able to obtain predictions or forecasts of future values.

3.3 The Lag and Differencing Operators

The *lag operator* L is defined for a time series $\{Y_t\}$ by

$$LY_t = Y_{t-1}.$$

The operator can be defined for linear combinations by

$$L(c_1Y_{t_1} + c_2Y_{t_2}) = c_1Y_{t_1-1} + c_2Y_{t_2-1}$$

and can be extended to all of \mathcal{H} by a suitable definition for limits.

In addition to being linear, the lag operator preserves inner products.

$$\begin{aligned} \langle LY_s, LY_t \rangle &= \text{cov}(Y_{s-1}, Y_{t-1}) \\ &= \text{cov}(Y_s, Y_t) \\ &= \langle Y_s, Y_t \rangle \end{aligned}$$

(An operator of this type is called a *unitary operator*.)

There is a natural calculus of operators on \mathcal{H} . For example we can define powers of L naturally by

$$\begin{aligned} L^2Y_t &= LLY_t = LY_{t-1} = Y_{t-2} \\ L^3Y_t &= LL^2Y_t = Y_{t-3} \\ &\vdots \\ L^kY_t &= Y_{t-k} \end{aligned}$$

and linear combinations by

$$(\alpha L^k + \beta L^l)Y_t = \alpha Y_{t-k} + \beta Y_{t-l}.$$

Other operators can be defined in terms in terms of L . The *differencing operator* defined by

$$\nabla Y_t = (1 - L)Y_t = Y_t - Y_{t-1}$$

is of fundamental importance when dealing with models for non-stationary time series. Again, we can define powers of this operator

$$\begin{aligned} \nabla^2Y_t &= \nabla(\nabla Y_t) \\ &= \nabla(Y_t - Y_{t-1}) \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2}. \end{aligned}$$

We will not dwell on the rich Hilbert space theory associated with time series, but it is important to know that many of the operator manipulations which we will carry out can be placed on a rigorous footing.

3.4 Linear Processes

We will now turn to an examination of a large class of useful time series models. These are almost all defined in terms of the lag operator L . As the simplest example, consider the autoregressive model defined by:

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \quad (3.1)$$

where ϕ is a constant with $|\phi| < 1$ and ε_t is a sequence of uncorrelated random variables, each with mean 0 and variance σ^2 . From a statistical point of view this “model” makes perfect sense, but is not clear that any Y_t which satisfies this equation exists.

One way to see that there is a solution is to re-arrange equation 3.1 and write it in its operator form.

$$(1 - \phi L)Y_t = \varepsilon_t.$$

Formally inverting the operator $(1 - \phi L)$ leads to

$$\begin{aligned} Y_t &= (1 - \phi L)^{-1} \varepsilon_t \\ &= \sum_{u=0}^{\infty} \phi^u L^u \varepsilon_t \\ &= \sum_{u=0}^{\infty} \phi^u \varepsilon_{t-u}. \end{aligned}$$

The series on the right is defined as the limit as $n \rightarrow \infty$ of

$$\sum_{u=0}^n \phi^u \varepsilon_{t-u}.$$

Loosely speaking, this limit exists if

$$\left\| \sum_{u=n+1}^{\infty} \phi^u \varepsilon_{t-u} \right\|^2 \rightarrow 0.$$

Since

$$\begin{aligned} \left\| \sum_{u=n+1}^{\infty} \phi^u \varepsilon_{t-u} \right\|^2 &= \text{var} \left(\sum_{u=n+1}^{\infty} \phi^u \varepsilon_{t-u} \right) \\ &= \sum_{u=n+1}^{\infty} |\phi|^{2u} \sigma^2 \end{aligned}$$

and $|\phi| < 1$, there is indeed a well-defined solution of 3.1. Further, this solution can be written as an (infinite) moving average of current and earlier ε_t values.

This type of infinite moving average plays a special role in the theory of time series.

Definition 3.4.1 (Linear Processes) The time series Y_t defined by

$$Y_t = \sum_{u=-\infty}^{\infty} \psi_u \varepsilon_{t-u}$$

where ε_t is a white-noise series and

$$\sum_{u=-\infty}^{\infty} |\psi_u|^2 < \infty$$

is called a *linear process*.

The general linear process depends on both past and future values of ε_t . A linear process which depends only on the past and present values of ε_t is said to be *causal*. Causal processes are preferred for forecasting because they reflect the way in which we believe the real world works.

Many time series can be represented as linear processes. This provides a unifying underpinning for time series theory, but may be of limited practical interest because of the potentially infinite number of parameters required.

3.5 Autoregressive Series

Definition 3.5.1 (Autoregressive Series) If Y_t satisfies

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

where ε_t is white-noise and the ϕ_u are constants, then Y_t is called an *autoregressive series of order p* , denoted by AR(p).

Autoregressive series are important because:

1. They have a natural interpretation — the next value observed is a slight perturbation of a simple function of the most recent observations.
2. It is easy to estimate their parameters. It can be done with standard regression software.
3. They are easy to forecast. Again standard regression software will do the job.

3.5.1 The AR(1) Series

The AR(1) series is defined by

$$Y_t = \phi Y_{t-1} + \varepsilon_t. \quad (3.2)$$

Because Y_{t-1} and ε_t are uncorrelated, the variance of this series is

$$\text{var}(Y_t) = \phi^2 \text{var}(Y_{t-1}) + \sigma_\varepsilon^2.$$

If $\{Y_t\}$ is stationary then $\text{var}(Y_t) = \text{var}(Y_{t-1}) = \sigma_Y^2$ and so

$$\sigma_Y^2 = \phi^2 \sigma_Y^2 + \sigma_\varepsilon^2. \quad (3.3)$$

This implies that

$$\sigma_Y^2 > \phi^2 \sigma_Y^2$$

and hence

$$1 > \phi^2.$$

There is an alternative view of this, using the operator formulation of equation 3.2, namely

$$(1 - \phi L)Y_t = \varepsilon_t$$

It is possible to formally invert the autoregressive operator to obtain

$$(1 - \phi L)^{-1} = \sum_{u=0}^{\infty} \phi^u L^u.$$

Applying this to the series $\{\varepsilon_t\}$ produces the representation

$$Y_t = \sum_{u=0}^{\infty} \phi^u \varepsilon_{t-u}. \quad (3.4)$$

If $|\phi| < 1$ this series converges in mean square because $\sum |\phi|^{2u} < \infty$. The limit series is stationary and satisfies equation 3.2. Thus, if $|\phi| < 1$ then there is a stationary solution to 3.2. An equivalent condition is that the root of the equation

$$1 - \phi z = 0$$

(namely $1/\phi$) lies outside the unit circle in the complex plane.

If $|\phi| > 1$ the series defined by 3.4 does not converge. We can, however, rearrange the defining equation 3.2 in the form

$$Y_t = \frac{1}{\phi} Y_{t+1} - \frac{1}{\phi} \varepsilon_{t+1},$$

and a similar argument to that above produces the representation

$$Y_t = - \sum_{u=0}^{\infty} \phi^{-u} \varepsilon_{t+u}.$$

This series does converge, so there is a stationary series Y_t which satisfies 3.2. The resulting series is generally not regarded as satisfactory for modelling and forecasting because it is not causal, i.e. it depends on future values of ε_t .

If $|\phi| = 1$ there is no stationary solution to 3.2. This means that for the purposes of modelling and forecasting stationary time series, we must restrict our attention to series for which $|\phi| < 1$ or, equivalently, to series for which the root of the polynomial $1 - \phi z$ lies outside the unit circle in the complex plane.

If we multiply both sides of equation 3.2 by Y_{t-u} and take expectations we obtain

$$E(Y_t Y_{t-u}) = \phi E(Y_{t-1} Y_{t-u}) + E(\varepsilon_t Y_{t-u}).$$

The term on the right is zero because, from the linear process representation, ε_t is independent of earlier Y_t values. This means that the autocovariances must satisfy the recursion

$$\gamma(u) = \phi \gamma(u-1), \quad u = 1, 2, 3, \dots$$

This is a first-order linear difference equation with solution

$$\gamma(u) = \phi^u \gamma(0), \quad u = 0, 1, 2, \dots$$

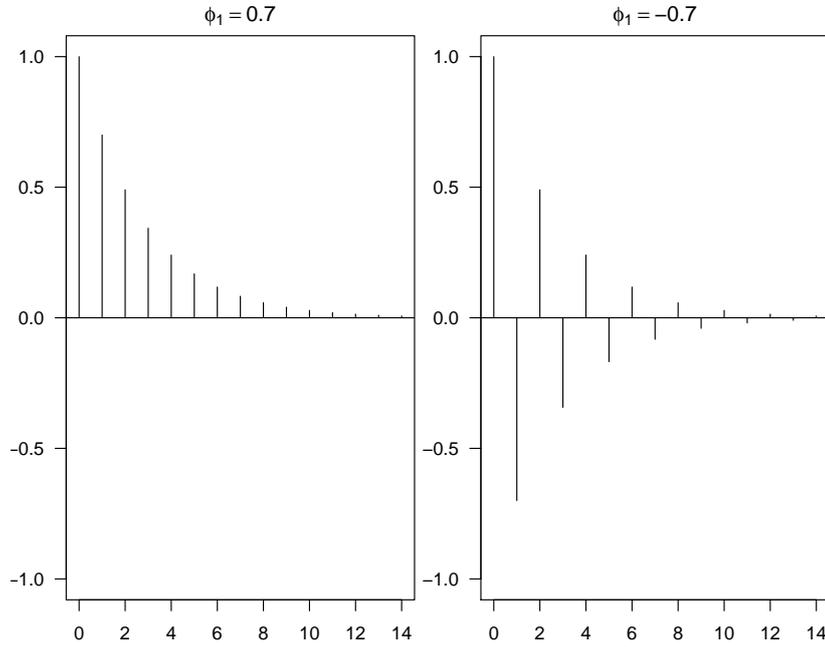


Figure 3.1: Autocorrelation functions for two of AR(1) models.

By rearranging equation 3.3 we find $\gamma(0) = \sigma_\varepsilon^2 / (1 - \phi^2)$, and hence that

$$\gamma(u) = \frac{\phi^u \sigma_\varepsilon^2}{1 - \phi^2}, \quad u = 0, 1, 2, \dots$$

This in turn means that the autocorrelation function is given by

$$\rho(u) = \phi^u, \quad u = 0, 1, 2, \dots$$

The autocorrelation functions for the the AR(1) series with $\phi_1 = .7$ and $\phi_1 = -.7$ are shown in figure 3.1. Both functions show exponential decay.

3.5.2 The AR(2) Series

The AR(2) model is defined by

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t \quad (3.5)$$

or, in operator form

$$(1 - \phi_1 L - \phi_2 L^2) Y_t = \varepsilon_t.$$

As in the AR(1) case we can consider inverting the AR operator. To see whether this is possible, we can consider factorising the operator

$$1 - \phi_1 L - \phi_2 L^2$$

and inverting each factor separately. Suppose that

$$1 - \phi_1 L - \phi_2 L^2 = (1 - c_1 L)(1 - c_2 L),$$

then it is clear that we can invert the operator if we can invert each factor separately. This is possible if $|c_1| < 1$ and $|c_2| < 1$, or equivalently, if the roots of the polynomial

$$1 - \phi_1 z - \phi_2 z^2$$

lie outside the unit circle. A little algebraic manipulation shows that this is equivalent to the conditions:

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

These constraints define a triangular region in the ϕ_1, ϕ_2 plane. The region is shown as the shaded triangle in figure 3.3.

The autocovariance function for the AR(2) series can be investigated by multiplying both sides of equation 3.5 by Y_{t-u} and taking expectations.

$$E(Y_t Y_{t-u}) = \phi_1 E(Y_{t-1} Y_{t-u}) + \phi_2 E(Y_{t-2} Y_{t-u}) + E(\varepsilon_t Y_{t-u}).$$

This in turns leads to the recurrence

$$\gamma(u) = \phi_1 \gamma(u-1) + \phi_2 \gamma(u-2)$$

with initial conditions

$$\begin{aligned} \gamma(0) &= \phi_1 \gamma(-1) + \phi_2 \gamma(-2) + \sigma_\varepsilon^2 \\ \gamma(1) &= \phi_1 \gamma(0) + \phi_2 \gamma(-1). \end{aligned}$$

or, using the fact that $\gamma(-u) = \gamma(u)$,

$$\begin{aligned} \gamma(0) &= \phi_1 \gamma(1) + \phi_2 \gamma(2) + \sigma_\varepsilon^2 \\ \gamma(1) &= \phi_1 \gamma(0) + \phi_2 \gamma(1). \end{aligned}$$

The solution to these equations has the form

$$\gamma(u) = A_1 G_1^u + A_2 G_2^u$$

where G_1^{-1} and G_2^{-1} are the roots of the polynomial

$$1 - \phi_1 z - \phi_2 z^2 \tag{3.6}$$

and A_1 and A_2 are constants that can be determined from the initial conditions. In the case that the roots are equal, the solution has the general form

$$\gamma(u) = (A_1 + A_2 u) G^u$$

These equations indicate that the autocovariance function for the AR(2) series will exhibit (exponential) decay as $u \rightarrow \infty$.

If G_k corresponds to a complex root, then

$$G_k = |G_k| e^{i\theta_k}$$

and hence

$$G_k^u = |G_k|^u e^{i\theta_k u} = |G_k|^u (\cos \theta_k u + i \sin \theta_k u)$$

Complex roots will thus introduce a pattern of decaying sinusoidal variation into the covariance function (or autocorrelation function). The region of the ϕ_1, ϕ_2 plane corresponding to complex roots is indicated by the cross-hatched region in figure 3.3.

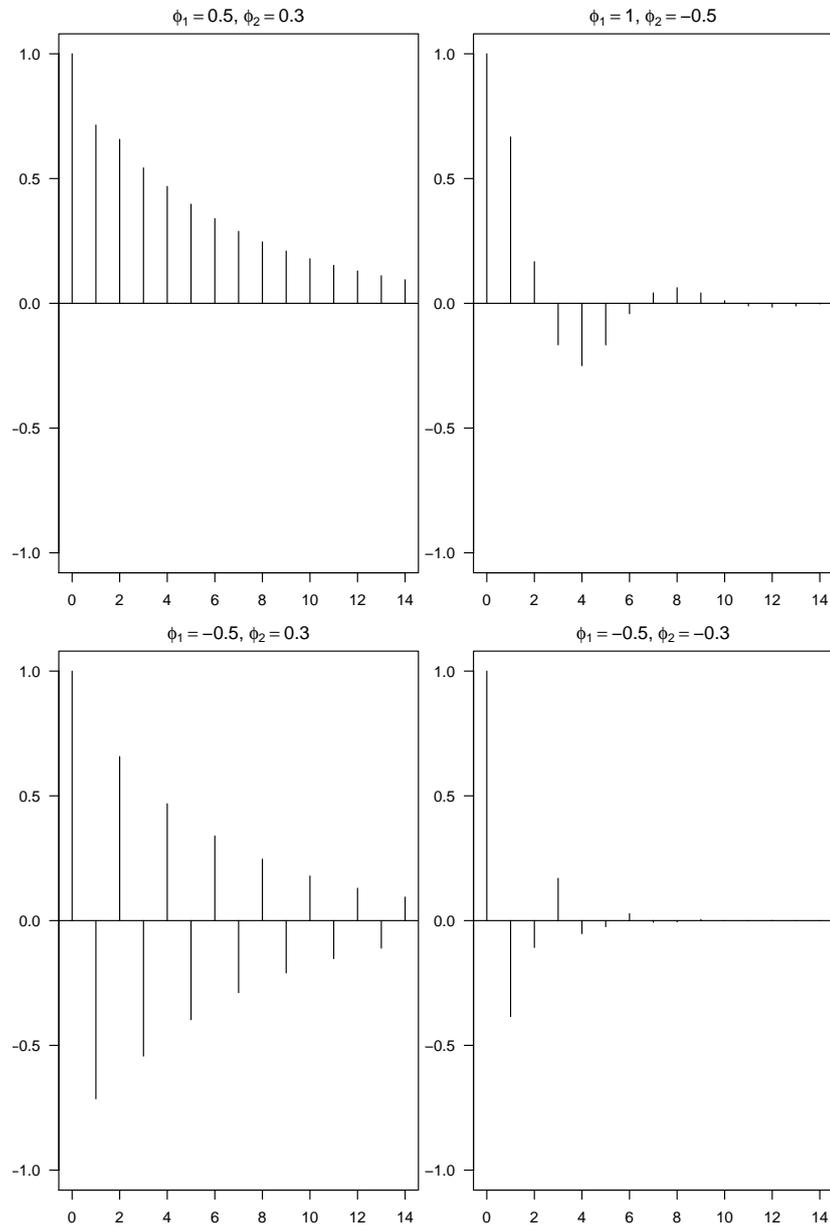


Figure 3.2: Autocorrelation functions for a variety of AR(2) models.

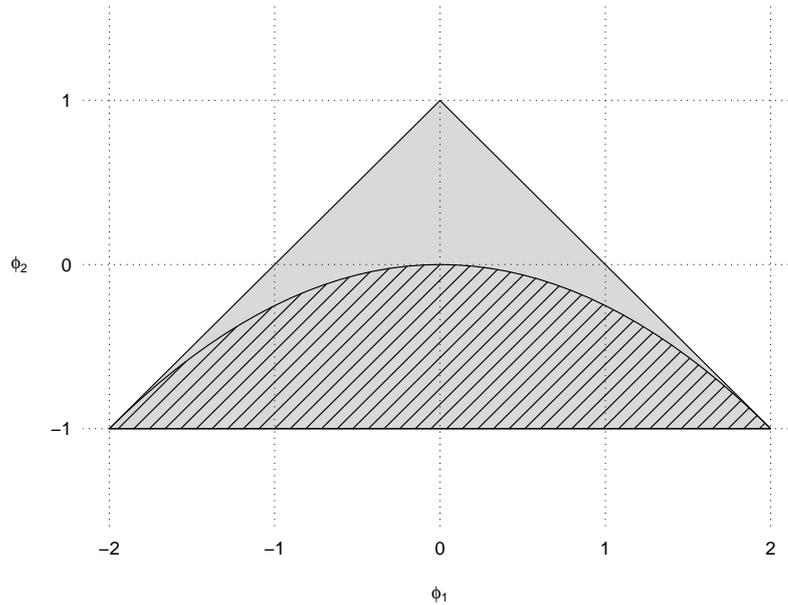


Figure 3.3: The regions of ϕ_1/ϕ_2 space where the series produced by the AR(2) scheme is stationary (indicated in grey) and has complex roots (indicated by cross-hatching).

The AR(p) Series

The AR(p) series is defined by

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \quad (3.7)$$

If the roots of

$$1 - \phi_1 z - \cdots - \phi_p z^p \quad (3.8)$$

lie outside the unit circle, there is a stationary causal solution to 3.7.

The autocovariance function can be investigated by multiplying equation 3.7 by Y_{t-u} and taking expectations. This yields

$$\gamma(u) = \phi_1 \gamma(u-1) + \cdots + \phi_p \gamma(u-p).$$

This is a linear homogeneous difference equation and has the general solution

$$\gamma(u) = A_1 G_1^u + \cdots + A_p G_p^u$$

(this is for distinct roots), where G_1, \dots, G_p are the reciprocals of the roots of equation 3.8. Note that the stationarity condition means that $\gamma(u) \rightarrow 0$, exhibiting exponential decay. As in the AR(2) case, complex roots will introduce oscillatory behaviour into the autocovariance function.

3.5.3 Computations

R contains a good deal of time series functionality. On older versions of R (those prior to 1.7.0) you will need to type the command

```
> library(ts)
```

to ensure that this functionality is loaded.

The function `polyroot` can be used to find the roots of polynomials and so determine whether a proposed model is stationary. Consider the model

$$Y_t = 2.5Y_{t-1} - Y_{t-2} + \varepsilon_t$$

or, its equivalent operator form

$$(1 - 2.5L + L^2)Y_t = \varepsilon_t.$$

We can compute magnitudes of the roots of the polynomial $1 - 2.5z + z^2$ with `polyroot`.

```
> Mod(polyroot(c(1,-2.5,1)))
[1] 0.5 2.0
```

The roots have magnitudes .5 and 2. Because the first of these is less than 1 in magnitude the model is thus not stationary and causal.

For the model

$$Y_t = 1.5Y_{t-1} - .75Y_{t-2} + \varepsilon_t$$

or its operator equivalent

$$(1 - 1.5L + .75L^2)Y_t = \varepsilon_t$$

we can check stationarity by examining the magnitudes of the roots of $1 - 1.5z + .75z^2$.

```
> Mod(polyroot(c(1,-1.5,.75)))
[1] 1.154701 1.154701
```

Both roots are bigger than 1 in magnitude, so the series is stationary. We can obtain the roots themselves as follows.

```
> polyroot(c(1,-1.5,.75))
[1] 1+0.5773503i 1-0.5773503i
```

Because the roots are complex we can expect to see a cosine-like ripple in the autocovariance function and the autocorrelation function.

The autocorrelation function for a given model can be computed using the `ARMAacf` function. The acf for the model above can be computed and plotted as follows.

```
> plot(0:14, ARMAacf(ar=c(1.5,-.75), lag=14), type="h",
      xlab = "Lag", ylab = "ACF")
> abline(h = 0)
```

The result is shown in figure 3.4.

Finally, it may be useful to simulate a time series of a given form. We can create a time series from the model $Y_t = 1.5Y_{t-1} - .75Y_{t-2} + \varepsilon_t$ and plot it with the following statements.

```
> x = arima.sim(model = list(ar=c(1.5,-.75)), n = 100)
> plot(x)
```

The result is shown in figure 3.5. Note that there is evidence that the series contains a quasi-periodic component with period about 12, as suggested by the autocorrelation function.

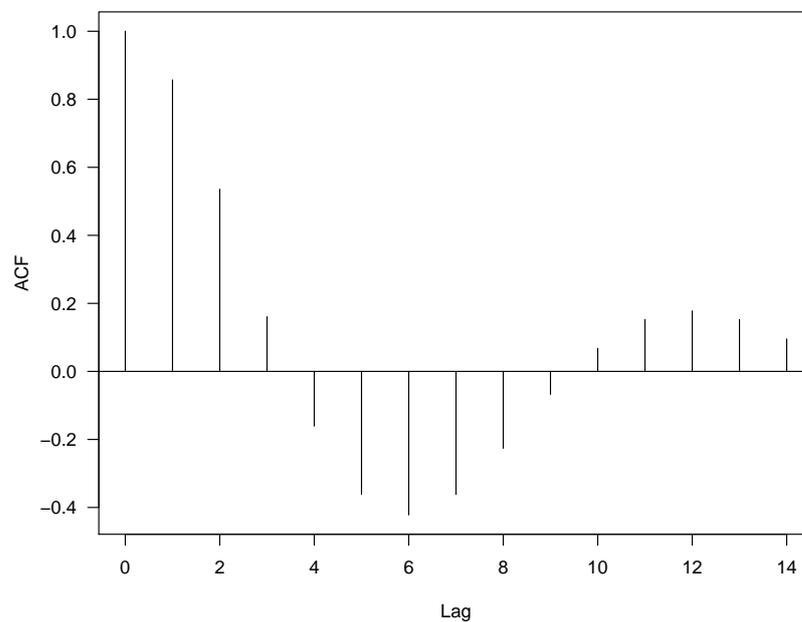


Figure 3.4: The acf for the model $Y_t = 1.5Y_{t-1} - .75Y_{t-2} + \varepsilon_t$.

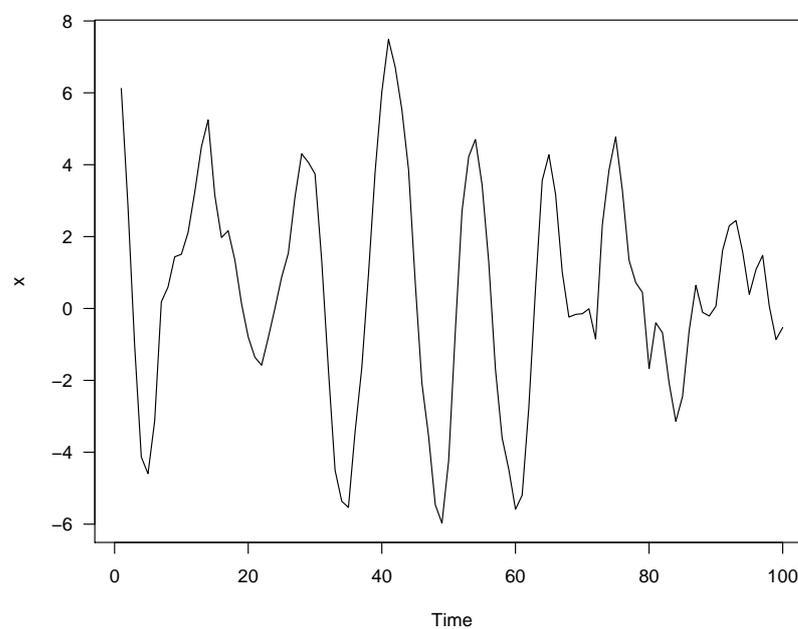


Figure 3.5: Simulation of the model $Y_t = 1.5Y_{t-1} - .75Y_{t-2} + \varepsilon_t$.

3.6 Moving Average Series

A time series $\{Y_t\}$ which satisfies

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (3.9)$$

(with $\{\varepsilon_t\}$ white noise) is said to be a moving average process of order q or MA(q) process. No additional conditions are required to ensure stationarity.

The autocovariance function for the MA(q) process is

$$\gamma(u) = \begin{cases} (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma^2 & u = 0 \\ (\theta_u + \theta_1\theta_{u+1} + \cdots + \theta_{q-u}\theta_q)\sigma^2 & u = 1, \dots, q \\ 0 & \text{otherwise.} \end{cases}$$

which says there is only a finite span of dependence on the series.

Note that it is easy to distinguish MA and AR series by the behaviour of their autocorrelation functions. The acf for MA series “cuts off” sharply while that for an AR series decays exponentially (with a possible sinusoidal ripple superimposed).

3.6.1 The MA(1) Series

The MA(1) series is defined by

$$Y_t = \varepsilon_t + \theta \varepsilon_{t-1}. \quad (3.10)$$

It has autocovariance function

$$\gamma(u) = \begin{cases} (1 + \theta^2)\sigma^2 & u = 0 \\ \theta\sigma^2 & u = 1 \\ 0 & \text{otherwise.} \end{cases}$$

and autocorrelation function

$$\rho(u) = \begin{cases} \frac{\theta}{1 + \theta^2} & \text{for } u = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (3.11)$$

3.6.2 Invertibility

If we replace θ by $1/\theta$ and σ^2 by $\theta\sigma^2$ the autocorrelation function given by 3.11 is unchanged. There are thus two sets of parameter values which can explain the structure of the series.

For the general process defined by equation 3.9, there is a similar identifiability problem. The problem can be resolved by requiring that the operator

$$1 + \theta_1 L + \cdots + \theta_q L^q$$

be invertible – i.e. that all roots of the polynomial

$$1 + \theta_1 z + \cdots + \theta_q z^q$$

lie outside the unit circle.

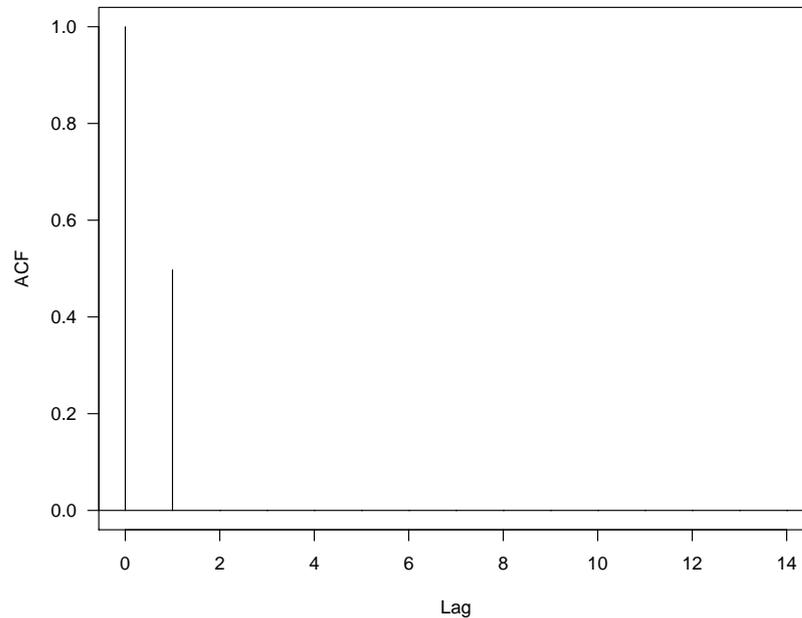


Figure 3.6: The acf for the model $Y_t = \varepsilon_t + .9\varepsilon_{t-1}$.

3.6.3 Computation

The function `polyroot` can be used to check invertibility for MA models. Remember that the invertibility requirement is only so that each MA model is only defined by one set of parameters.

The function `ARMAacf` can be used to compute the acf for MA series. For example, the acf of the model

$$Y_t = \varepsilon_t + 0.9\varepsilon_{t-1}$$

can be computed and plotted as follows.

```
> plot(0:14, ARMAacf(ma=.9, lag=14), type="h",
      xlab = "Lag", ylab = "ACF")
> abline(h = 0)
```

The result is shown in figure 3.6.

A simulation of the series can be computed and plotted as follows.

```
> x = arima.sim(model = list(ma=.9), n = 100)
> plot(x)
```

The result of the simulation is shown in figure 3.7.

3.7 Autoregressive Moving Average Series

Definition 3.7.1 If a series satisfies

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (3.12)$$

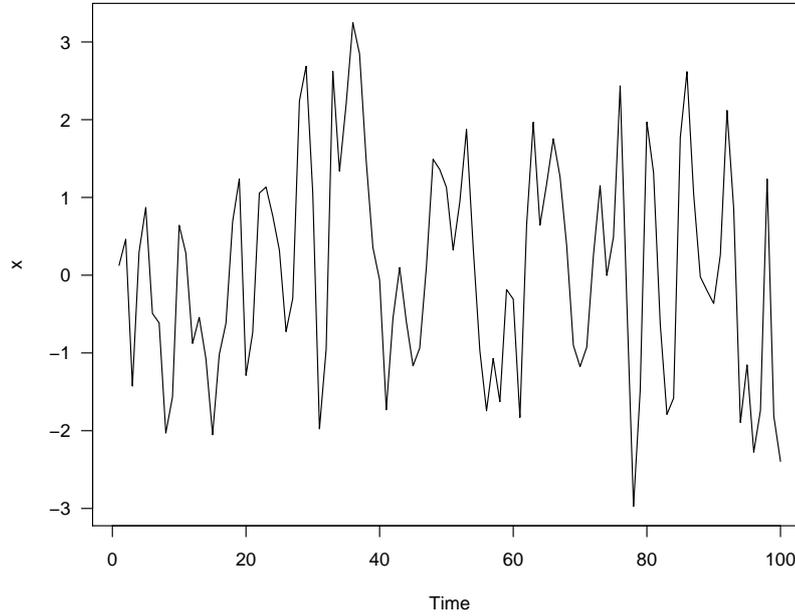


Figure 3.7: Simulation of the model $Y_t = \varepsilon_t + .9\varepsilon_{t-1}$.

(with $\{\varepsilon_t\}$ white noise), it is called an *autoregressive-moving average series of order (p, q)* , or an $\text{ARMA}(p, q)$ series.

An $\text{ARMA}(p, q)$ series is stationary if the roots of the polynomial

$$1 - \phi_1 z - \cdots - \phi_p z^p$$

lie outside the unit circle.

3.7.1 The ARMA(1,1) Series

The $\text{ARMA}(1,1)$ series is defined by

$$Y_t = \phi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}. \quad (3.13)$$

To derive the autocovariance function for Y_t , note that

$$\begin{aligned} E(\varepsilon_t Y_t) &= E[\varepsilon_t(\phi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1})] \\ &= \sigma_\varepsilon^2 \end{aligned}$$

and

$$\begin{aligned} E(\varepsilon_{t-1} Y_t) &= E[\varepsilon_{t-1}(\phi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1})] \\ &= \phi \sigma_\varepsilon^2 + \theta \sigma_\varepsilon^2 \\ &= (\phi + \theta) \sigma_\varepsilon^2. \end{aligned}$$

Multiplying equation 3.13 by Y_{t-u} and taking expectation yields:

$$\gamma(u) = \begin{cases} \phi\gamma(1) + (1 + \theta(\phi + \theta))\sigma_\varepsilon^2 & u = 0 \\ \phi\gamma(0) + \theta\sigma_\varepsilon^2 & u = 1 \\ \phi\gamma(u-1) & u \geq 2 \end{cases}$$

Solving the first two equations produces

$$\gamma(0) = \frac{(1 + 2\theta\phi + \theta^2)}{1 - \phi^2}\sigma_\varepsilon^2 = \frac{(1 + 2\theta\phi + \theta^2)}{1 - \phi^2}\sigma_\varepsilon^2$$

and using the last recursively shows

$$\gamma(u) = \frac{(1 + \theta\phi)(\phi + \theta)}{1 - \phi^2}\phi^{u-1}\sigma_\varepsilon^2 \quad \text{for } u \geq 1.$$

The autocorrelation function can then be computed as

$$\rho(u) = \frac{(1 + \theta\phi)(\phi + \theta)}{(1 + 2\theta\phi + \theta^2)}\phi^{u-1} \quad \text{for } u \geq 1.$$

The pattern here is similar to that for AR(1), except for the first term.

3.7.2 The ARMA(p, q) Model

It is possible to make general statements about the behaviour of general ARMA(p, q) series. When values are more than q time units apart, the memory of the moving-average part of the series is lost. The functions $\gamma(u)$ and $\rho(u)$ will then behave very similarly to those for the AR(p) series

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \varepsilon_t$$

for large u , but the first few terms will exhibit additional structure.

3.7.3 Computation

Stationarity can be checked by examining the roots of the characteristic polynomial of the AR operator and model parameterisation can be checked by examining the roots of the characteristic polynomial of the MA operator. Both checks can be carried out with `polyroot`.

The autocorrelation function for an ARMA series can be computed with `ARMAacf`. For the model

$$Y_t = -.5Y_{t-1} + \varepsilon_t + .3\varepsilon_{t-1}$$

this can be done as follows

```
> plot(0:14, ARMAacf(ar=-.5, ma=.3, lag=14), type="h",
      xlab = "Lag", ylab = "ACF")
> abline(h = 0)
```

and produces the result shown in figure 3.8.

Simulation can be carried out using `arma.sim`

```
> x = arima.sim(model = list(ar=-.5, ma=.3), n = 100)
> plot(x)
```

producing the result in figure 3.9.

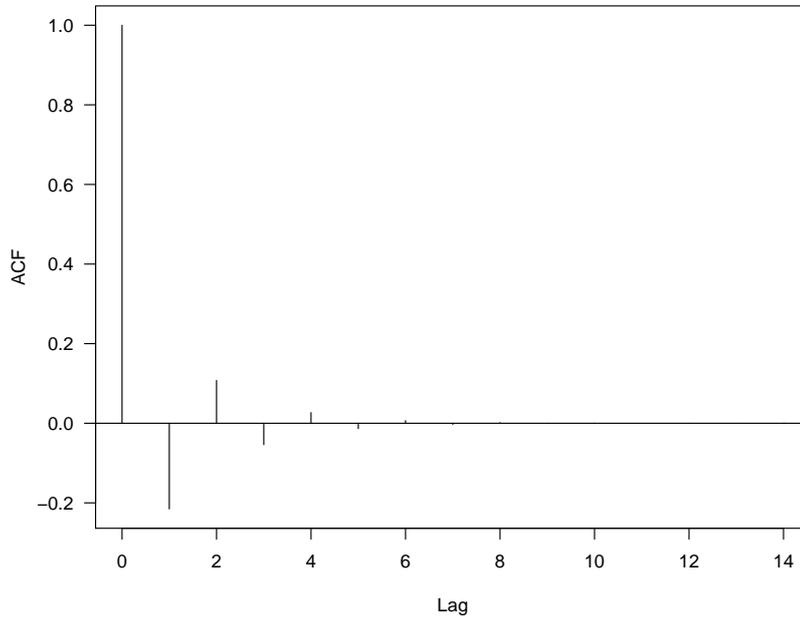


Figure 3.8: The acf for the model $Y_t = -0.5Y_{t-1} + \epsilon_t + 0.3\epsilon_{t-1}$.

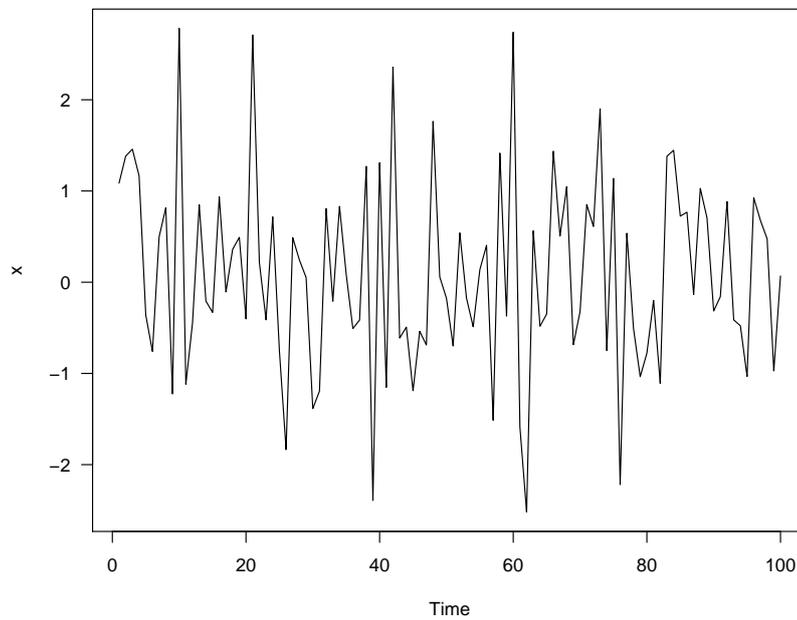


Figure 3.9: Simulation of the model $Y_t = -0.5Y_{t-1} + \epsilon_t + 0.3\epsilon_{t-1}$.

3.7.4 Common Factors

If the AR and MA operators in an ARMA(p, q) model possess common factors, then the model is over-parameterised. By dividing through by the common factors we can obtain a simpler model giving an identical description of the series. It is important to recognise common factors in an ARMA model because they will produce numerical problems in model fitting.

3.8 The Partial Autocorrelation Function

The autocorrelation function of an MA series exhibits different behaviour from that of AR and general ARMA series. The acf of an MA series cuts off sharply whereas those for AR and ARMA series exhibit exponential decay (with possible sinusoidal behaviour superimposed). This makes it possible to identify an ARMA series as being a purely MA one just by plotting its autocorrelation function. The partial autocorrelation function provides a similar way of identifying a series as a purely AR one.

Given a stretch of time series values

$$\dots, Y_{t-u}, Y_{t-u+1}, \dots, Y_{t-1}, Y_t, \dots$$

the partial correlation of Y_t and Y_{t-u} is the correlation between these random variables which is not conveyed through the intervening values.

If the Y values are normally distributed, the partial autocorrelation between Y_t and Y_{t-u} can be defined as

$$\phi(u) = \text{cor}(Y_t, Y_{t-u} | Y_{t-1}, \dots, Y_{t-u+1}).$$

A more general approach is based on regression theory. Consider predicting Y_t based on $Y_{t-1}, \dots, Y_{t-u+1}$. The prediction is

$$\hat{Y}_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} \dots, \beta_{u-1} Y_{t-u+1}$$

with the β s chosen to minimise

$$E(Y_t - \hat{Y}_t)^2.$$

It is also possible to “think backwards in time” and consider predicting Y_{t-u} with the same set of predictors. The best predictor will be

$$\hat{Y}_{t-u} = \beta_1 Y_{t-u+1} + \beta_2 Y_{t-u+2} \dots, \beta_{u-1} Y_{t-1}.$$

(The coefficients are the same because the correlation structure is the same whether the series is run forwards or backwards in time.)

The partial correlation function at lag u is the correlation between the prediction errors.

$$\phi(u) = \text{cor}(Y_t - \hat{Y}_t, Y_{t-u} - \hat{Y}_{t-u})$$

By convention we take $\phi(1) = \rho(1)$.

It is quite straightforward to compute the value of $\phi(2)$. Using the results of section 2.4.1, the best predictor of Y_t based on Y_{t-1} is just $\rho(1)Y_{t-1}$. Thus

$$\begin{aligned}\text{cov}(Y_t - \rho(1)Y_{t-1}, Y_{t-2} - \rho(1)Y_{t-1}) &= \sigma_Y^2(\rho(2) - \rho(1)^2 - \rho(1)^2) + \rho(1)^2 \\ &= \sigma_Y^2(\rho(2) - \rho(1)^2)\end{aligned}$$

and

$$\begin{aligned}\text{var}(Y_t - \rho(1)Y_{t-1}) &= \sigma_Y^2(1 + \rho(1)^2 - 2\rho(1)^2) \\ &= \sigma_Y^2(1 - \rho(1)^2)^2\end{aligned}$$

This means that

$$\phi(2) = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2} \quad (3.14)$$

Example 3.8.1 For the AR(1) series, recall that

$$\rho(u) = \phi^u \quad (u \geq 0).$$

Substituting this into equation 3.14 we find

$$\phi(2) = \frac{\phi^2 - \phi^2}{1 - \phi^2} = 0.$$

Example 3.8.2 For the MA(1) series

$$\rho(u) = \begin{cases} \frac{\theta}{1 + \theta^2} & \text{if } u = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Substituting this into 3.14 we find

$$\begin{aligned}\phi(2) &= \frac{0 - (\theta/(1 + \theta^2))^2}{1 - (\theta/(1 + \theta^2))^2} \\ &= \frac{-\theta^2}{(1 + \theta^2)^2 - \theta^2} \\ &= \frac{-\theta^2}{1 + \theta^2 + \theta^4}.\end{aligned}$$

More generally it is possible to show

$$\phi(u) = \frac{-\theta^u(1 - \theta^2)}{1 - \theta^{2(u+1)}} \quad \text{for } u \geq 0.$$

For the general AR(p) series, it is possible to show that $\phi(u) = 0$ for all $u > p$. For such a series, the best predictor of Y_t using $Y_{t-1}, \dots, Y_{t-u+1}$ for $u > p$ is

$$\phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p}.$$

because

$$Y_t - \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} = \varepsilon_t$$

and ε_t is uncorrelated with Y_{t-1}, Y_{t-2}, \dots , so that the “fit” cannot be improved.

The prediction error corresponding to the best linear predictor of Y_{t-u} is based on $Y_{t-1}, \dots, Y_{t-u+1}$ and so must be uncorrelated with ε_t . This shows that $\phi(u) = 0$.

For the general MA(q), it is possible to show that $\phi(u)$ decays exponentially as $u \rightarrow \infty$.

3.8.1 Computing the PACF

The definition of the partial autocorrelation function given in the previous section is conceptually simple, but it makes computations hard. In this section we’ll see that there is an equivalent form which is computationally simple.

Consider the k th order autoregressive prediction of Y_{k+1}

$$\widehat{Y}_{k+1} = \phi_{k1} Y_k + \cdots + \phi_{kk} Y_1 \quad (3.15)$$

obtained by minimising $E(Y_{k+1} - \widehat{Y}_{k+1})^2$. We will show that the k th partial autocorrelation values is given by $\phi(k) = \phi_{kk}$. The proof of this is a geometric one which takes places in the space \mathcal{H} generated by the series $\{Y_t\}$.

We begin by defining the subspace $\mathcal{H}_1 = \overline{\text{sp}}\{Y_2, \dots, Y_k\}$ and associated projection $\mathcal{P}_{\mathcal{H}_1}$, the subspace $\mathcal{H}_2 = \overline{\text{sp}}\{Y_1 - \mathcal{P}_{\mathcal{H}_1} Y_1\}$ and the subspace $\mathcal{H}_k = \overline{\text{sp}}\{Y_1, \dots, Y_k\}$. Any $Y \in \mathcal{H}$,

$$\mathcal{P}_{\mathcal{H}_k} Y = \mathcal{P}_{\mathcal{H}_1} Y + \mathcal{P}_{\mathcal{H}_2} Y.$$

Thus

$$\begin{aligned} \widehat{Y}_{k+1} &= \mathcal{P}_{\mathcal{H}_k} Y_{k+1} \\ &= \mathcal{P}_{\mathcal{H}_1} Y_{k+1} + \mathcal{P}_{\mathcal{H}_2} Y_{k+1} \\ &= \mathcal{P}_{\mathcal{H}_1} Y_{k+1} + a(Y_1 - \mathcal{P}_{\mathcal{H}_1} Y_1), \end{aligned}$$

where

$$a = \langle Y_{k+1}, Y_1 - \mathcal{P}_{\mathcal{H}_1} Y_1 \rangle / \|Y_1 - \mathcal{P}_{\mathcal{H}_1} Y_1\|^2. \quad (3.16)$$

Rearranging, we find

$$\widehat{Y}_{k+1} = \mathcal{P}_{\mathcal{H}_1}(Y_{k+1} - aY_1) + aY_1.$$

The first term on the right must be a linear combination of Y_2, \dots, Y_k , so comparing with equation 3.15 we see that $a = \phi_{kk}$.

Now, the k th partial correlation is defined as the correlation between the residuals from the regressions of Y_{k+1} and Y_1 on Y_2, \dots, Y_k . But this is just

$$\begin{aligned} \text{cor}(Y_{k+1} - \mathcal{P}_{\mathcal{H}_1} Y_{k+1}, Y_1 - \mathcal{P}_{\mathcal{H}_1} Y_1) \\ &= \langle Y_{k+1} - \mathcal{P}_{\mathcal{H}_1} Y_{k+1}, Y_1 - \mathcal{P}_{\mathcal{H}_1} Y_1 \rangle / \|Y_1 - \mathcal{P}_{\mathcal{H}_1} Y_1\|^2 \\ &= \langle Y_{k+1}, Y_1 - \mathcal{P}_{\mathcal{H}_1} Y_1 \rangle / \|Y_1 - \mathcal{P}_{\mathcal{H}_1} Y_1\|^2 \\ &= a \end{aligned}$$

by equation 3.16.

A recursive way of computing the regression coefficients in equation 3.15 from the autocorrelation function was given by Levinson (1947) and Durbin (1960). The Durbin-Levinson algorithm updates the coefficients the from $k - 1$ st order model to those of the k th order model as follows:

$$\phi_{kk} = \frac{\rho(k) - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_j}$$

$$\phi_{k,j} = \phi_{k-1,j} - \phi_{kk} \phi_{k-1,k-j} \quad j = 1, 2, \dots, k - 1.$$

3.8.2 Computation

The R function `ARMAacf` can be used to obtain the partial autocorrelation function associated with a stationary ARMA series. The call to `ARMAacf` is identical to its use for obtaining the ordinary autocorrelation function, except it has the additional argument `pacf=TRUE`.

The following code computes and plots the partial autocorrelation function for the ARMA(1,1) model with $\phi = -.5$ and $\theta = .3$.

```
> plot(1:14, ARMAacf(ar=-.5, ma=.3, lag=14, pacf=TRUE),
      type="h", xlab = "Lag", ylab = "ACF")
> abline(h = 0)
```

The resulting plot is shown in figure 3.10.

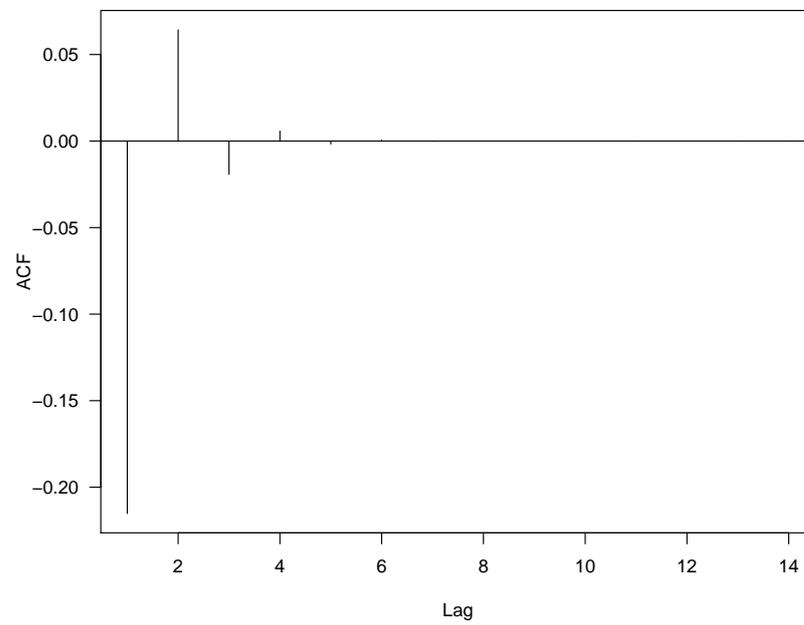


Figure 3.10: The partial acf for the model $Y_t = -0.5Y_{t-1} + \epsilon_t + 0.3\epsilon_{t-1}$.

Chapter 4

Identifying Time Series Models

4.1 ACF Estimation

We have seen that it is possible to distinguish between AR, MA and ARMA models by the behaviour of their acf and pacf functions. In practise, we don't know these functions and so we must estimate them.

Given a stretch of data Y_1, \dots, Y_T , the usual estimate of the autocovariance function is

$$\hat{\gamma}(u) = \frac{1}{T} \sum_{t=1}^{T-u} (Y_{t+u} - \bar{Y})(Y_t - \bar{Y})$$

Note that this estimator is *biased* — an unbiased estimator would have a divisor of $T - u$ in place of T . There are two reasons for using this estimator.

The first of these reasons is that it produces a $\hat{\gamma}(u)$ which is *positive definite*. This means that for any constants c_1, \dots, c_k ,

$$\sum_{u=1}^k \sum_{v=1}^k c_u c_v \hat{\gamma}(u - v) \geq 0.$$

This ensures that our estimate of the variance of

$$\sum_{u=1}^k c_u X_{t-u}$$

will be non-negative, something which might not be the case for the unbiased estimate.

The second reason is that for many time series $\gamma(u) \rightarrow 0$ as $u \rightarrow \infty$. For such time series, the biased estimate can have lower mean-squared error.

The estimate of $\rho(u)$ based on $\hat{\gamma}(u)$ is

$$r(u) = \frac{\hat{\gamma}(u)}{\hat{\gamma}(0)} = \frac{\sum_t (Y_{t+u} - \bar{Y})(Y_t - \bar{Y})}{\sum_t (Y_t - \bar{Y})^2}$$

(Again, this can have better mean-squared error properties than the estimate based on the unbiased estimate of $\gamma(u)$.)

In order to say whether an observed correlation is significantly different from zero, we need some distribution theory. Like most time series results, the theory here is asymptotic (as $T \rightarrow \infty$). The original results in this area were obtained by Bartlett in 1947. We will look at results due to T. W. Anderson in 1971.

Suppose that

$$Y_t = \mu + \sum_{u=0}^{\infty} \psi_u \varepsilon_{t-u}$$

with the ε_t independent and identically distributed with zero mean and non-zero variance. Suppose that

$$\sum_{u=0}^{\infty} |\psi_u| < \infty \quad \text{and} \quad \sum_{u=0}^{\infty} u |\psi_u|^2 < \infty.$$

(This is true for all stationary ARMA series). The last condition can be replaced by the requirement that the $\{Y_t\}$ values have a finite fourth moment.

Under these conditions, for any fixed m , the joint distribution of

$$\sqrt{T}(r(1) - \rho(1)), \sqrt{T}(r(2) - \rho(2)), \dots, \sqrt{T}(r(m) - \rho(m))$$

is asymptotically normal with zero means and covariances

$$\begin{aligned} c_{uv} = & \sum_{t=0}^{\infty} (\rho(t+u)\rho(t+v) + \rho(t-u)\rho(t+v) \\ & - 2\rho(u)\rho(t)\rho(t+v) - 2\rho(v)\rho(t)\rho(t+u) \\ & + 2\rho(u)\rho(v)\rho(t)^2). \end{aligned} \quad (4.1)$$

I.e. for large T

$$r(u) \approx N(0, c_{uu}/T) \quad \text{cor}(r(u), r(v)) \approx \frac{c_{uv}}{\sqrt{c_{uu}c_{vv}}}.$$

Notice that $\text{var } r(u) \downarrow 0$ but that the correlations stay approximately constant.

Equation 4.1 is clearly not easy to interpret in general. Let's examine some special cases.

Example 4.1.1 White Noise.

The theory applies to the case that the Y_t are i.i.d.

$$\text{var } r(u) \approx \frac{1}{T} \quad \text{cor}(r(u), r(v)) \approx 0$$

Example 4.1.2 The AR(1) Series.

In this case $\rho(u) = \phi^u$ for $u > 0$. After a good deal of algebra (summing geometric series) one finds:

$$\text{var } r(u) \approx \frac{1}{T} \left(\frac{(1 + \phi^2)(1 - \phi^{2u})}{1 - \phi^2} - 2u\phi^{2u} \right).$$

In particular, for $u = 1$,

$$\text{var } r(1) \approx \frac{1 - \phi^2}{T}.$$

Table 4.1: Large Sample Results for r_k for an AR(1) Model.

ϕ	$\sqrt{\text{var } r(1)}$	$\sqrt{\text{var } r(2)}$	$\text{cor}(r(1), r(2))$	$\sqrt{\text{var } r(10)}$
0.9	$0.44/\sqrt{T}$	$0.807/\sqrt{T}$	0.97	$2.44/\sqrt{T}$
0.7	$0.71/\sqrt{T}$	$1.12/\sqrt{T}$	0.89	$1.70/\sqrt{T}$
0.5	$0.87/\sqrt{T}$	$1.15/\sqrt{T}$	0.76	$1.29/\sqrt{T}$
0.3	$0.95/\sqrt{T}$	$1.08/\sqrt{T}$	0.53	$1.09/\sqrt{T}$
0.1	$0.99/\sqrt{T}$	$1.01/\sqrt{T}$	0.20	$1.01/\sqrt{T}$

Notice that the closer ϕ is to ± 1 , the more accurate the estimate becomes.

As $u \rightarrow \infty$, $\phi^{2u} \rightarrow 0$. In that case

$$\text{var } r(u) \approx \frac{1}{T} \left(\frac{1 + \phi^2}{1 - \phi^2} \right).$$

For values of ϕ close to ± 1 this produces large variances for the $r(u)$.

For $0 < u \leq v$ (after much algebra),

$$c_{uv} = \frac{(\phi^{v-1} - \phi^{v+u})(1 + \phi^2)}{1 - \phi^2} + (v - u)\phi^{v-u} - (v + u)\phi^{v+u}.$$

In particular,

$$\text{cor}(r(1), r(2)) \approx 2\phi \left(\frac{1 - \phi^2}{1 + 2\phi^2 - 3\phi^4} \right)^{1/2}$$

Using these formulae it is possible to produce the results in table 4.1.

Example 4.1.3 The MA(1) Series

For the MA(1) series it is straightforward to show that

$$c_{11} = 1 - 3\rho(1)^2 + 4\rho(1)^4$$

$$c_{uu} = 1 - 2\rho(1)^2 \quad u > 1$$

$$c_{12} = 2\rho(1)(1 - \rho(1)^2)$$

Using these results it is easy to produce the results in table 4.2.

Example 4.1.4 The General MA(q) Series

In the case of the general MA(q) series it is easy to see that

$$c_{uu} = 1 + 2 \sum_{v=1}^q \rho(v)^2, \quad \text{for } u > q,$$

and hence that

$$\text{var } r(u) = \frac{1}{T} \left(1 + 2 \sum_{v=1}^q \rho(v)^2 \right), \quad \text{for } u > q.$$

Table 4.2: Large Sample Results for r_k for an MA(1) Model.

θ	$\sqrt{\text{var } r(1)}$	$\sqrt{\text{var } r(k)} (k > 1)$	$\text{cor}(r(1), r(2))$
0.9	$0.71/\sqrt{T}$	$1.22/\sqrt{T}$	0.86
0.7	$0.73/\sqrt{T}$	$1.20/\sqrt{T}$	0.84
0.5	$0.79/\sqrt{T}$	$1.15/\sqrt{T}$	0.74
0.4	$0.84/\sqrt{T}$	$1.11/\sqrt{T}$	0.65

Notes

1. In practise we don't know the parameters of the model generating the data we might have. We can still estimate the variances and covariances of the $r(u)$ by substituting estimates of $\rho(u)$ into the formulae above.
2. Note that there can be quite large correlations between the $r(u)$ values so caution must be used when examining plots of $r(u)$.

4.2 PACF Estimation

In section 3.8.1 we saw that the theoretical pacf can be computed by solving the Durbin-Levinson recursion

$$\phi_{kk} = \frac{\rho(k) - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho(k-j)}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho(j)}$$

$$\phi_{k,j} = \phi_{k-1,j} - \phi_{kk} \phi_{k-1,k-j} \quad j = 1, 2, \dots, k-1.$$

and setting $\phi(u) = \phi_{uu}$.

In practice, the estimated autocorrelation function is used in place of the theoretical autocorrelation function to generate estimates of the partial autocorrelation function.

To decide whether partial autocorrelation values are significantly different from zero, we can use a (1949) result of Quenouille which states that if the true underlying model is AR(p), then the estimated partial autocorrelations at lags greater than p are approximately independently normal with means equal to zero and variance $1/T$. Thus $\pm 2/\sqrt{T}$ can be used as critical limits on $\phi(u)$ for $u > p$ to test the hypothesis of an AR(p) model.

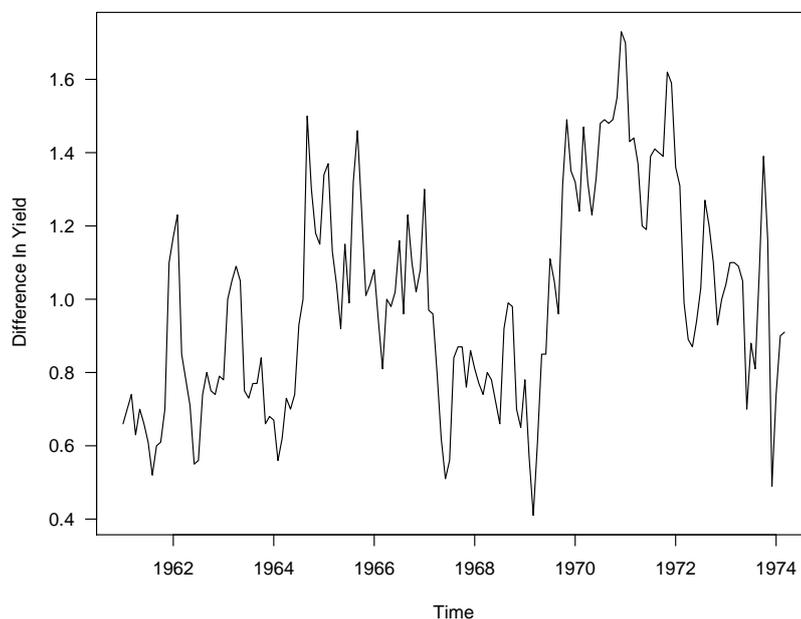


Figure 4.1: Monthly differences between the yield on mortgages and government loans in the Netherlands, January 1961 to March 1974.

4.3 System Identification

Given a set of observations Y_1, \dots, Y_T we will need to decide what the appropriate model might be. The estimated acf and pacf are the tools which can be used to do this. If the acf exhibits slow decay and the pacf cuts off sharply after lag p , we would identify the series as $AR(p)$. If the pacf shows slow decay and the acf show a sharp cutoff after lag q , we would identify the series as being $MA(q)$. If both the acf and pacf show slow decay we would identify the series as being mixed ARMA. In this case the orders of the AR and MA parts are not clear, but it is reasonable to first try $ARMA(1,1)$ and move on to higher order models if the fit of this model is not good.

Example 4.3.1 Interest Yields

Figure 4.1 shows a plot of the Monthly differences between the yield on mortgages and government loans in the Netherlands, January 1961 to March 1974. The series appears stationary, so we can attempt to use the acf and pacf to decide whether an AR, MA or ARMA model might be appropriate.

Figures 4.2 and 4.3 show the estimated acf and pacf functions for the yield series. The horizontal lines in the plots are drawn at the y values $\pm 1.96/\sqrt{159}$ (the series has 159 values). These provide 95% confidence limits for what can be expected under the hypothesis of white noise. Note that these limits are *point-wise* so that we would expect to see roughly 5% of the values lying outside the limits.

The acf plot shows evidence of slow decay, while the pacf plot shows a “sharp

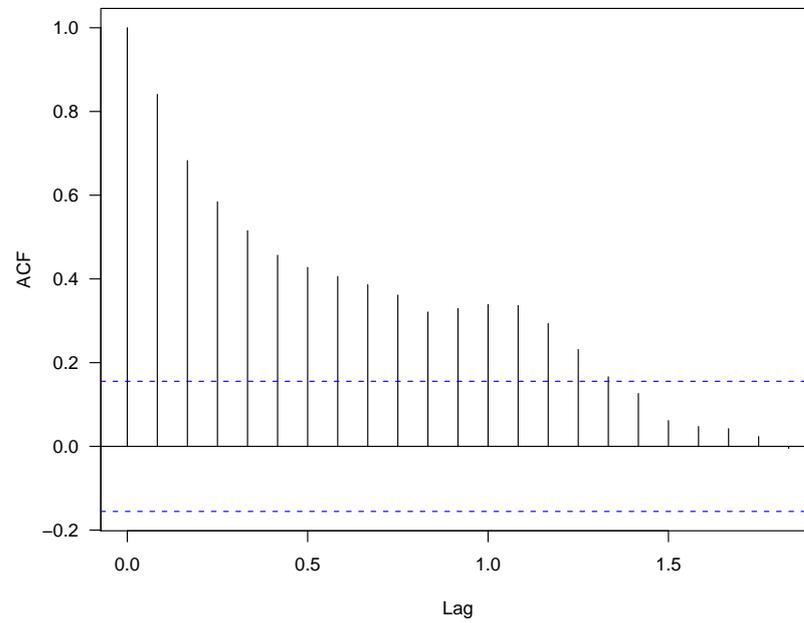


Figure 4.2: The acf for the yield data.

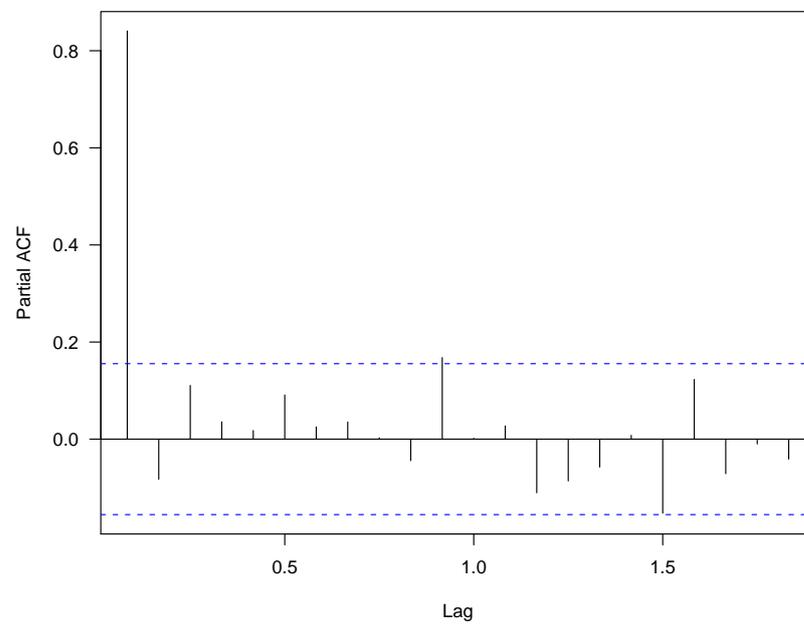


Figure 4.3: The pacf for the yield data.

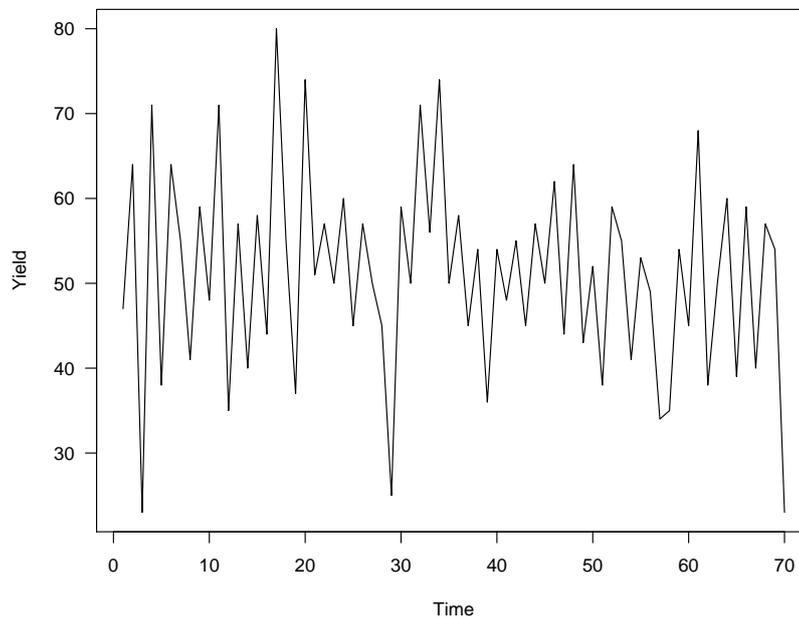


Figure 4.4: Yields from a chemical process (from Box and Jenkins).

cutoff” after lag 1. On the basis of these two plots we might hypothesise that an AR(1) model was an appropriate description of the data.

Example 4.3.2 Box and Jenkins Chemical Yields

Figure 4.4 shows an example from the classic time series text by Box and Jenkins. This series contains consecutive yields recorded from a chemical process. Again, the series is apparently stationary so that we can consider identifying an appropriate model on the basis of the acf and pacf.

Again, the acf seems to show slow decay, this time with alternating signs. The pacf shows sudden cutoff after lag 1, suggesting that again an AR(1) model might be appropriate.

4.4 Model Generalisation

ARMA series provide a flexible class of models for stationary mean-zero series, with AR and MA series being special cases.

Unfortunately, many series are clearly not in this general class for models. It is worthwhile looking at some generalisation of this class.

4.4.1 Non-Zero Means

When a series $\{Y_t\}$ has a non-zero mean μ , the mean can be subtracted and the deviations from the mean modeled as an ARMA series.

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}$$

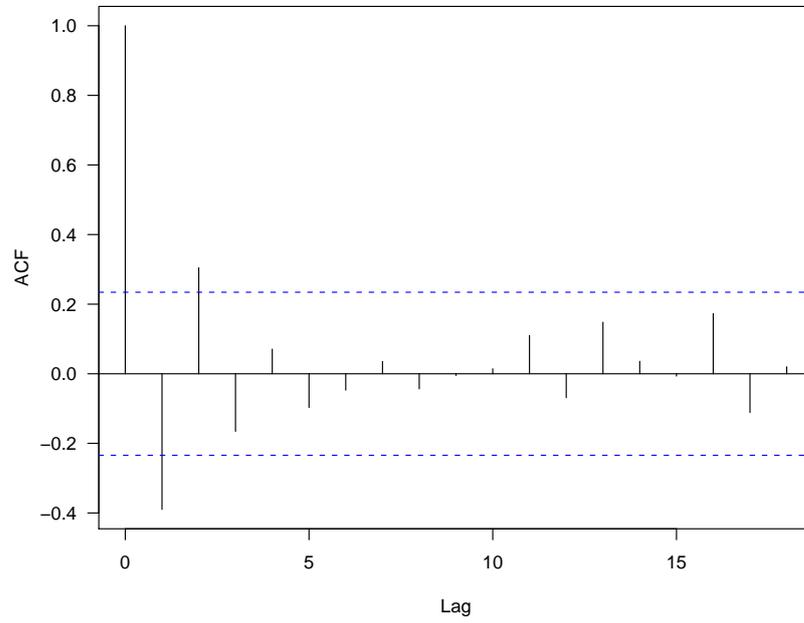


Figure 4.5: The acf for the chemical process data.

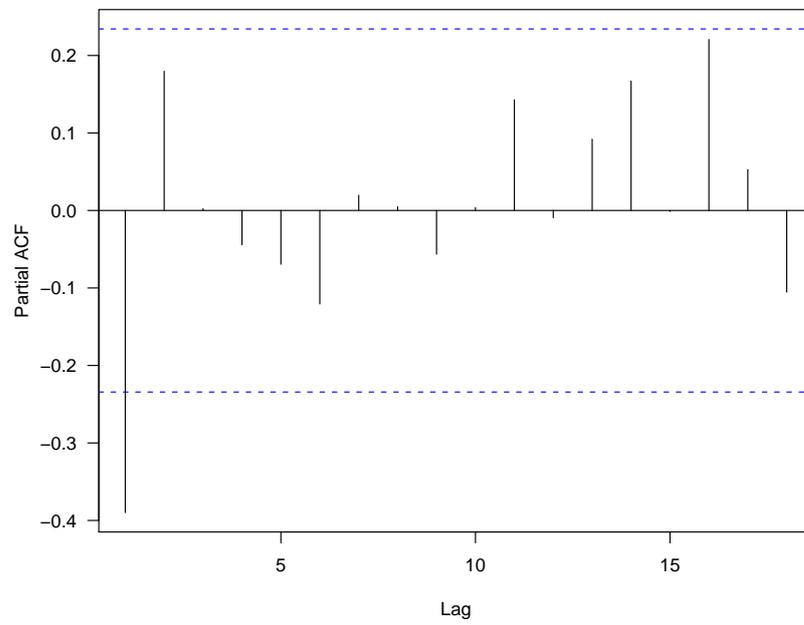


Figure 4.6: The pacf for the chemical process data.

Alternatively, the model can be adjusted by introducing a constant directly.

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

The two characterisations are connected by

$$\theta_0 = \mu - \mu(\phi_1 + \cdots + \phi_p)$$

so that

$$\mu = \frac{\theta_0}{1 - \phi_1 - \cdots - \phi_p}$$

or

$$\theta_0 = \mu(1 - \phi_1 - \cdots - \phi_p).$$

4.4.2 Deterministic Trends

Consider the model

$$Y_t = f(t) + Z_t,$$

where Z_t is a stationary ARMA series and $f(t)$ is a deterministic function of t . Considering $Y_t - f(t)$ reduces Y_t to an ARMA series. (If $f(t)$ contains unknown parameters we can estimate them.)

4.4.3 Models With Non-stationary AR Components

We've seen that any AR model with characteristic equation roots outside the unit circle will be non-stationary.

Example 4.4.1 Random Walks

A random walk is defined by the equation

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

where $\{\varepsilon_t\}$ is a series of uncorrelated (perhaps independent) random variables. In operator form this equation is

$$(1 - L)Y_t = \varepsilon_t.$$

The equation $1 - z = 0$ has a root at 1 so that Y_t is non-stationary.

Example 4.4.2 Integrated Moving Averages

Consider the model

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

This is similar to an ARMA(1,1) model, but again it is non-stationary.

Both these models can be transformed to stationarity by differencing — transforming to $\nabla Y_t = Y_t - Y_{t-1}$. In the first example we have,

$$\nabla Y_t = \varepsilon_t,$$

which is the white noise model. In the second we have,

$$\nabla Y_t = \varepsilon_t + \theta \varepsilon_{t-1},$$

which is the MA(1) model.

4.4.4 The Effect of Differencing

Suppose that Y_t has a linear trend

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

where Z_t is stationary with $E(Z_t) = 0$.

$$\nabla Y_t = \beta_1 + \nabla Z_t$$

Differencing has removed the trend. Now suppose that $\{Y_t\}$ has a deterministic quadratic trend

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + Z_t$$

then

$$\begin{aligned} \nabla Y_t &= (\beta_0 + \beta_1 t + \beta_2 t^2) - (\beta_0 + \beta_1(t-1) + \beta_2(t-1)^2) + \nabla Z_t \\ &= \beta_1 + \beta_2(t^2 - (t^2 - 2t + 1)) + \nabla Z_t \\ &= (\beta_1 - \beta_2) + 2\beta_2 t + \nabla Z_t \\ &= \text{linear trend} + \text{stationary}. \end{aligned}$$

Differencing again produces

$$\nabla^2 Y_t = 2\beta_2 + \nabla^2 Z_t.$$

In general, a polynomial trend of order k can be eliminated by differencing k times.

Now let's consider the case of a "stochastic trend." Suppose that

$$Y_t = M_t + \varepsilon_t$$

where M_t is a random process which changes slowly over time. In particular, we can assume that M_t is generated by a random walk model.

$$M_t = M_{t-1} + \eta_t$$

with η_t independent of ε_t . Then

$$\nabla Y_t = \eta_t + \varepsilon_t - \varepsilon_{t-1}.$$

∇Y_t is stationary and has an autocorrelation function like that of an MA(1) series.

$$\rho(1) = \frac{-1}{2 + (\sigma_\eta / \sigma_\varepsilon)^2}$$

More generally, "higher order" stochastic trend models can be reduced to stationarity by repeated differencing.

4.5 ARIMA Models

If $W_t = \nabla^d Y_t$ is an ARMA(p, q) series than Y_t is said to be an *integrated autoregressive moving-average* (p, d, q) series, denoted ARIMA(p, d, q). If we write

$$\phi(L) = 1 - \phi_1 L - \dots - \phi_p L^p$$

and

$$\theta(L) = 1 + \theta_1 L + \dots + \theta_q L^q$$

then we can write down the operator formulation

$$\phi(L)\nabla^d Y_t = \theta(L)\varepsilon_t.$$

Example 4.5.1 The IMA(1,1) Model This model is widely used in business and economics. It is defined by

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

Y_t can be thought of as a random walk with correlated errors.

Notice that

$$\begin{aligned} Y_t &= Y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1} \\ &= Y_{t-2} + \varepsilon_{t-1} + \theta\varepsilon_{t-2} + \varepsilon_t + \theta\varepsilon_{t-1} \\ &= Y_{t-2} + \varepsilon_t + (1 + \theta)\varepsilon_{t-1} + \theta\varepsilon_{t-2} \\ &\quad \vdots \\ &= Y_{-m} + \varepsilon_t + (1 + \theta)\varepsilon_{t-1} + \dots + (1 + \theta)\varepsilon_{-m} + \theta\varepsilon_{-m-1} \end{aligned}$$

If we assume that $Y_{-m} = 0$ (i.e. observation started at time $-m$),

$$Y_t = \varepsilon_t + (1 + \theta)\varepsilon_{t-1} + \dots + (1 + \theta)\varepsilon_{-m} + \theta\varepsilon_{-m-1}$$

This representation can be used to derive the formulae

$$\begin{aligned} \text{var}(Y_t) &= (1 + \theta^2 + (1 + \theta)^2(t + m))\sigma_\varepsilon^2 \\ \text{cor}(Y_k, Y_{t-k}) &= \frac{1 + \theta^2 + (1 + \theta)^2(t + m - k)}{(\text{var}(Y_t)\text{var}(Y_{t-k}))^{1/2}} \\ &= \sqrt{\frac{t + m - k}{t + m}} \\ &\approx 1 \end{aligned}$$

for m large and k moderate. (We are considering behaviour after “burn-in.”) This means that we can expect to see very slow decay in the autocorrelation function.

The very slow decay of the acf function is characteristic of ARIMA series with $d > 0$. Figures 4.7 and 4.8 show the estimated autocorrelation functions from a simulated random walk and an integrated autoregressive model. Both functions show very slow declines in the acf.

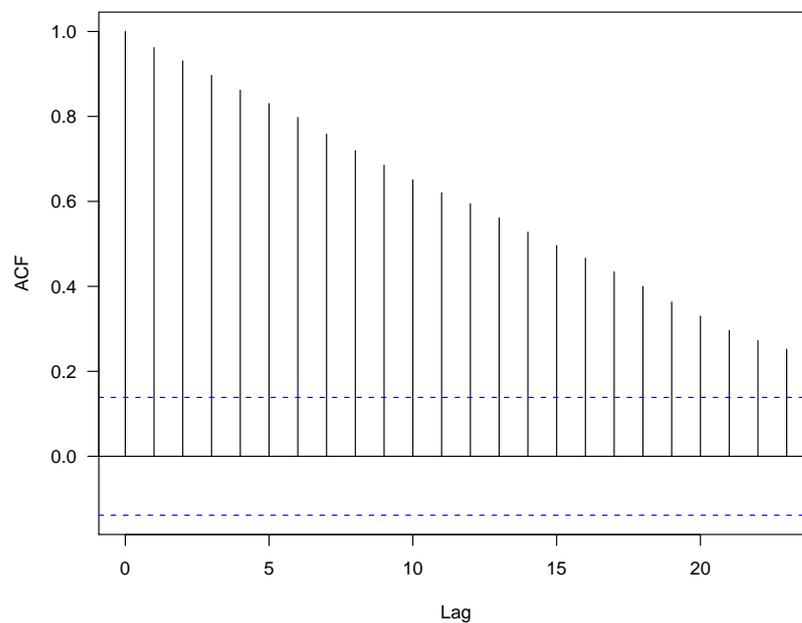


Figure 4.7: The autocorrelation function of the ARIMA(0,1,0) (random walk) model.

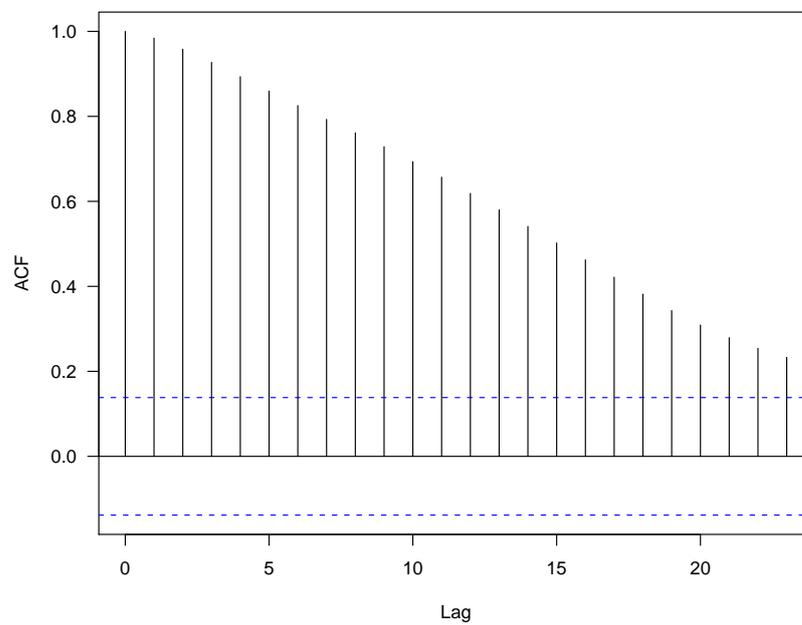


Figure 4.8: The autocorrelation function of the ARIMA(1,1,0) model with $\phi_1 = .5$.

Chapter 5

Fitting and Forecasting

5.1 Model Fitting

Suppose that we have identified a particular ARIMA(p, d, q) model which appears to describe a given time series. We now need to fit the identified model and assess how well the model fits. Fitting is usually carried out using maximum likelihood. For a given set of model parameters, we calculate a series of one-step-ahead predictions.

$$\hat{Y}_{k+1} = P_{\mathcal{H}_k} Y_{k+1}$$

where \mathcal{H}_k is the linear space spanned by Y_1, \dots, Y_k . The predictions are obtained in a recursive fashion using a process known as *Kalman filtering*. Each prediction results in a prediction error $\hat{Y}_{k+1} - Y_{k+1}$. These are, by construction, uncorrelated. If we add the requirement that the $\{Y_t\}$ series is normally distributed, the prediction errors are independent normal random variables and this can be used as the basis for computing the likelihood.

The parameters which need to be estimated are the AR coefficients ϕ_1, \dots, ϕ_p , the MA coefficients $\theta_1, \dots, \theta_q$ and a constant term (either μ or θ_0 as outlined in section 4.4.1). Applying maximum-likelihood produces both estimates and standard errors.

5.1.1 Computations

Given a time series y in R, we can fit an ARMA(p, d, q) model to the series as follows

```
> z = arima(y, order=c(p, d, q))
```

The estimation results can be inspected by printing them.

Example 5.1.1 U.S. Unemployment Rates

Figure 5.1 shows a plot of the seasonally adjusted quarterly United States unemployment rates from the first quarter of 1948 to the first quarter of 1978. If the data is stored as a time series in the R data set `unemp`, the plot can be produced with the command

```
> plot(unemp)
```

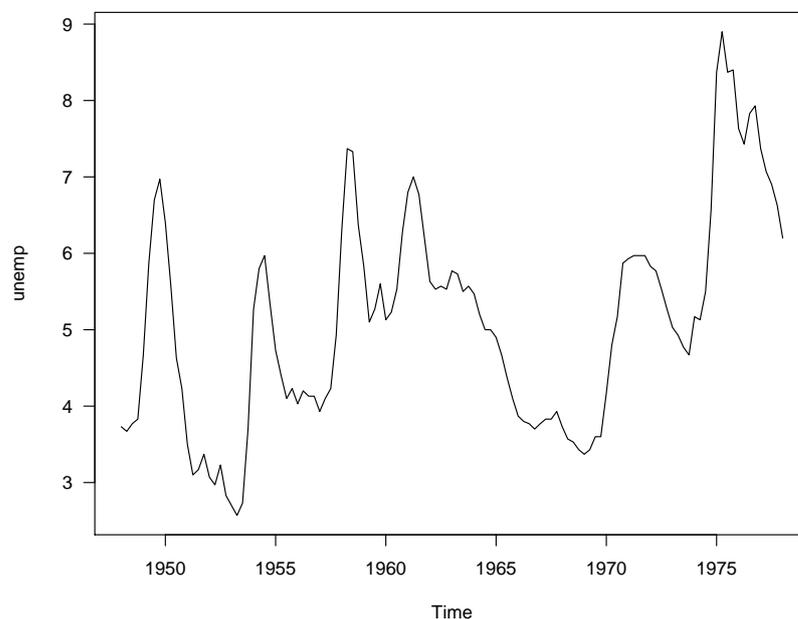


Figure 5.1: United States quarterly unemployment rates (seasonally adjusted).

Neither the original series nor the presence of very slow decay in the acf (figure 5.2) indicate strongly that differencing is required so we will leave the series undifferenced and attempt to find an appropriate ARMA model.

The acf and pacf functions for the series can be computed and plotted with the following commands

```
> acf(unemp)
> pacf(unemp)
```

The resulting plots (figures 5.2 and 5.3) show the strong signature of an AR(2) series (slow decay of the acf and sharp cutoff of the pacf after two lags.)

With the series identified we can go about estimating the unknown parameters. We do this with the `arima` function in R.

```
> z = arima(unemp, order=c(2, 0, 0))
> z
```

Call:

```
arima(x = unemp, order = c(2, 0, 0))
```

Coefficients:

	ar1	ar2	intercept
	1.5499	-0.6472	5.0815
s.e.	0.0681	0.0686	0.3269

```
sigma^2 estimated as 0.1276: log likelihood = -48.76, aic = 105.53
```

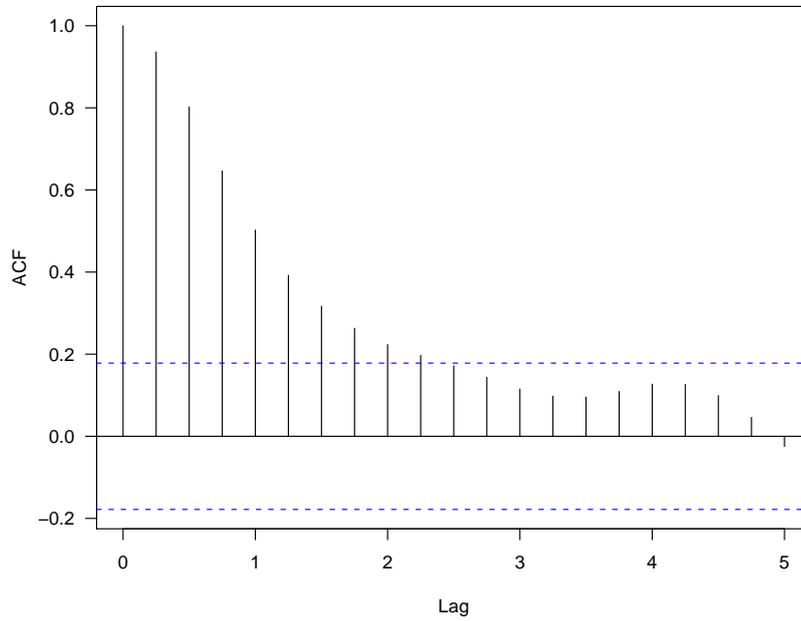


Figure 5.2: The acf for the unemployment series.

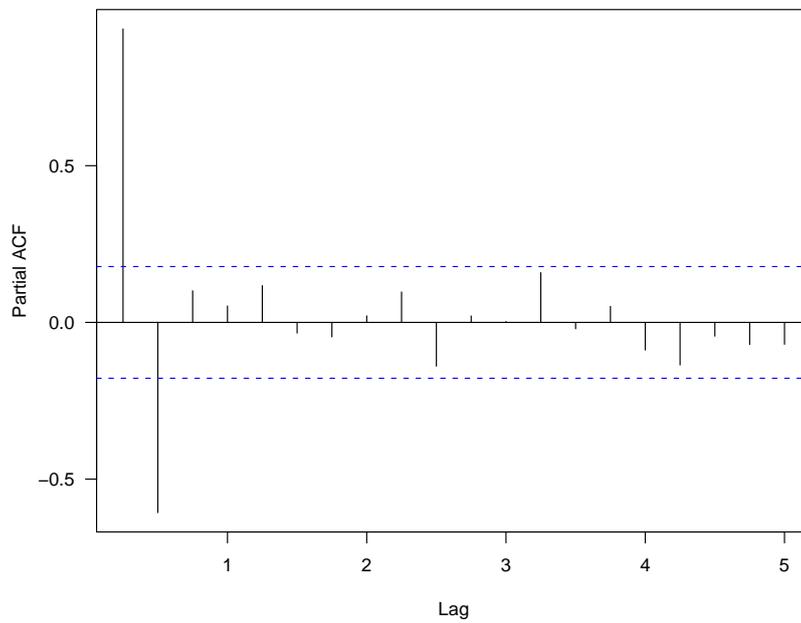


Figure 5.3: The pacf for the unemployment series.

The fitted model in this case is

$$Y_t = 5.0815 + 1.5499 Y_{t-1} - 0.6472 Y_{t-2} + \varepsilon_t$$

where ε_t has an estimated variance of 0.1276.

Example 5.1.2 Railroad Bond Yields

Figure 5.4 shows a plot of the monthly yields on AA rated railroad bonds (as a percentage times 100). With the data values stored in the R variable `rrbonds`, the plot was produced with the command

```
> plot(rrbonds, ylab = "Bond Yields")
```

In this case it is clear that the series is non-stationary and we need to transform to stationarity by taking differences. The series can be differenced and the result plotted as follows

```
> rrdiffs = diff(rrbonds)
> plot(rrdiffs, ylab = "Differenced Yields")
```

Figure 5.5 shows a plot of the differenced series and it is clear from this plot that the assumption of stationarity is much more reasonable. To confirm this, and to suggest possible models we need to examine the acf and pacf functions. These are shown in figures 5.6 and 5.7.

The model signature here is less certain than that of the previous example. A possible interpretation is that the acf is showing rapid decay and the pacf is showing sharp cutoff after one lag. This suggests that original series can be modelled as an ARIMA(1,1,0) series.

The estimation of parameters can be carried out for this model as follows

```
> z = arima(rrbonds, order=c(1,1,0))
> z
```

Call:

```
arima(x = rrbonds, order = c(1, 1, 0))
```

Coefficients:

```
      ar1
      0.4778
s.e.  0.0865
```

```
sigma^2 estimated as 84.46:  log likelihood = -367.47,  aic = 738.94
```

The fitted model in this case is

$$\nabla Y_t = 0.4778 \nabla Y_t + \varepsilon_t.$$

with ε_t having an estimated variance of 84.46.

5.2 Assessing Quality of Fit

Once a model has been fitted to a set of data it is always important to assess how well the model fits. This is because the inferences we make depend crucially



Figure 5.4: Monthly AA railroad bond yields ($\% \times 100$).

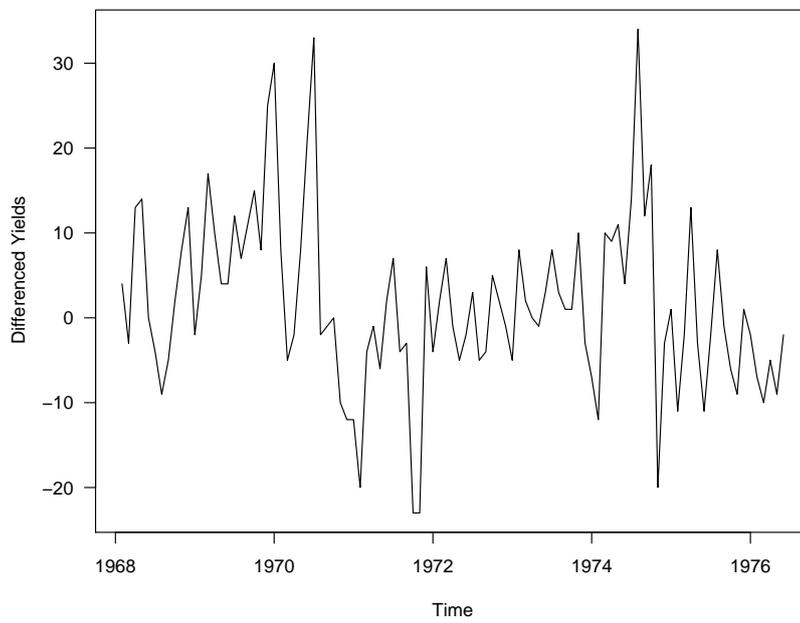


Figure 5.5: Differenced monthly AA railroad bond yields.

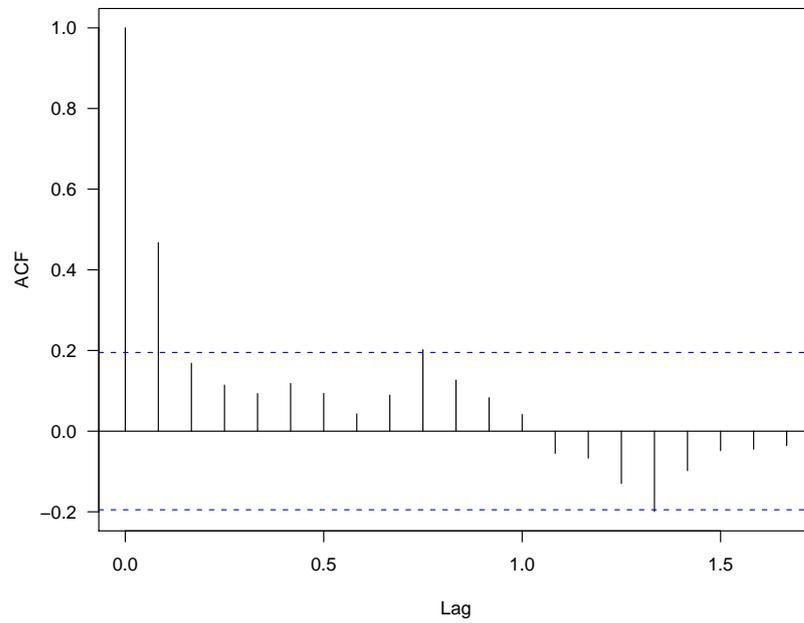


Figure 5.6: The ACF for the differenced monthly AA railroad bond yields.

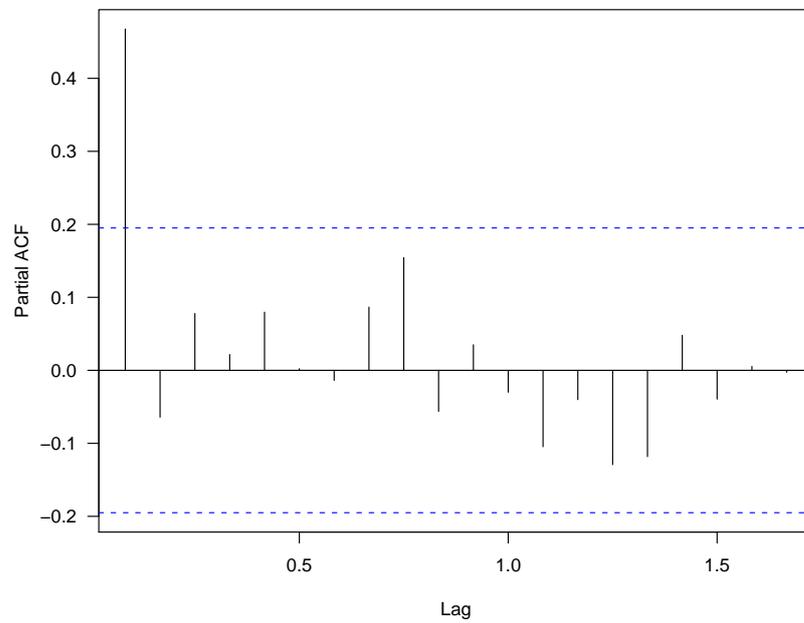


Figure 5.7: The PACF for the differenced monthly AA railroad bond yields.

on the appropriateness of the fitted model. The usual way of assessing goodness of fit is through the examination of residuals.

In the case of autoregressive models it is easy to see how residuals might be defined. In the case of the AR(p) model

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

it is clear that we can take the residuals to be

$$\hat{\varepsilon}_t = Y_t - \hat{\phi}_1 Y_{t-1} + \cdots + \hat{\phi}_p Y_{t-p}$$

In the general ARMA case

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

or

$$\phi(L) = \theta(L)\varepsilon_{t-q}$$

we must first transform to autoregressive form by inverting the MA operator.

$$\theta(L)^{-1}\phi(L) = \varepsilon_t$$

or

$$Y_t = \sum_{u=1}^{\infty} \pi_u Y_{t-u} + \varepsilon_t.$$

The residuals can then be defined as

$$\hat{\varepsilon}_t = Y_t - \sum_{u=1}^{\infty} \hat{\pi}_u Y_{t-u}.$$

This is a useful theoretical approach, but in practise the residuals are obtained as a byproduct of the computation of the likelihood (they are the prediction errors from the one-step ahead forecasts). If the ARMA model is correct (and the series is normally distributed) then the residuals are approximately independent normal random variables with mean zero and variance σ_ε^2 .

A simple diagnostic is to simply plot the residuals and to see whether they appear to be a white noise series. In the case of the US unemployment series we can do this as follows

```
> z = arima(unemp, order=c(2, 0, 0))
> plot(resid(z))
```

The results are shown in figure 5.8.

It is also possible to carry out tests of normality on the residuals. This can be done by simply producing a histogram of the residuals or by performing a normal quantile-quantile plot. This can be done as follows:

```
> hist(resid(z))
> qqnorm(resid(z))
```

A simple check of heteroscedasticity can be carried out by plotting the residuals against the fitted values. This is done with the following command:

```
> plot(unemp - resid(z), resid(z),
+      xy.lines = FALSE, xy.labels = FALSE)
```

None of these plots indicate that there is any problem with the fit of the model.

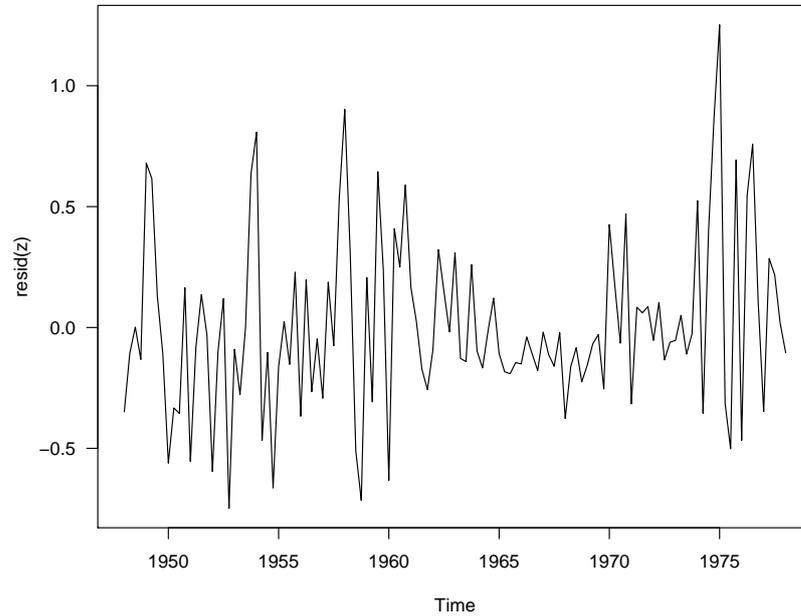


Figure 5.8: Residuals from the unemployment data.

5.3 Residual Correlations

It is tempting to examine the quality of model fit by seeing whether the residuals form an uncorrelated sequence. One might for example plot the estimated acf of the residuals and look for lags where the correlations exceed $\pm 2/\sqrt{T}$. Unfortunately, while these limits are approximately correct for large lags, for small lags they overstate the variability of the estimated correlations. This should be no surprise, the effect of model-fitting is to remove as much of the correlation present in the series as possible. The correlations between the residuals should be closer to zero than for a non-fitted series.

In addition to checking whether there are large individual correlations present in a time series, it can be useful to pool information from successive correlations to see whether there is significant correlation left in the residuals.

One test statistic in common use is the modified Box-Pierce (or Ljung-Box-Pierce) statistic

$$Q^* = T(T+2) \sum_{k=1}^K \frac{\widehat{r}_k^2}{T-k}.$$

If the true underlying model is $\text{ARMA}(p,q)$, the distribution of Q^* is approximately χ_{K-p-q}^2 .

In R, the modified Box-Pierce statistic (and an older variant) can be computed with the function `Box.test` (you should specify `type="Ljung"`). An alternative is to use the function `tsdiag` which plots the (standardised) residuals, the acf of the residuals and the modified Box-Pierce statistic for a variety of values of K . In the case of the US unemployment series, an appropriate

command would be

```
> tsdiag(z)
```

This produces the plot shown in figure 5.9. The first panel of the plot shows the (standardised) residuals from the model fit. The residuals look reasonably random but during the period from 1965 to 1970 there is a sequence of values which are all negative. This indicates that there may be problems with the model fit. The second panel shows the autocorrelation function for the residuals. One of the correlations (that at a lag of 7 quarters) lies outside the two standard error bounds about zero. Since we would expect one in twenty of the estimated correlations to exceed these bounds this does not provide evidence for a significant correlation. The third panel shows p-values for the Ljung-Box statistics at lags from one quarter to ten quarters. There is no evidence of significant correlation in this plot.

Overall, there is some evidence of lack of fit, but there is no obvious way of fixing the problem. Because of this we will move on to making forecasts for the series, but keep in mind that we need to be cautious in making use of these forecasts.

5.4 Forecasting

Once we have decided on an appropriate time-series model, estimated its unknown parameters and established that the model fits well, we can turn to the problem of forecasting future values of the series.

The autoregressive representation

$$Y_t = \sum_{u=1}^{\infty} \pi_u Y_{t-u} + \varepsilon_t.$$

suggests predicting the next observation beyond Y_1, \dots, Y_T using

$$\hat{Y}_{T+1} = \sum_{u=1}^{\infty} \hat{\pi}_u Y_{T+1-u}.$$

where the $\hat{\pi}_u$ are obtained by substituting the estimated parameters in place of the theoretical ones.

Once a forecast is obtained for Y_{T+1} we can use it to obtain a forecast for Y_{T+2} and then use these two forecasts to generate a forecast for Y_{T+3} . The process can be continued to obtain forecasts out to any point in the future. Because uncertainty increases as we predict further and further from the data we have, we can expect the standard errors associated with our predictions to increase.

In practise, forecasts are generated by the same Kalman filter algorithm used to compute the likelihood used for parameter estimation.

5.4.1 Computation

Once a model has been fitted using `arima`, the function `predict` can be used to obtain forecasts. In the case of the US unemployment series we could obtain 10 future forecasts as follows:

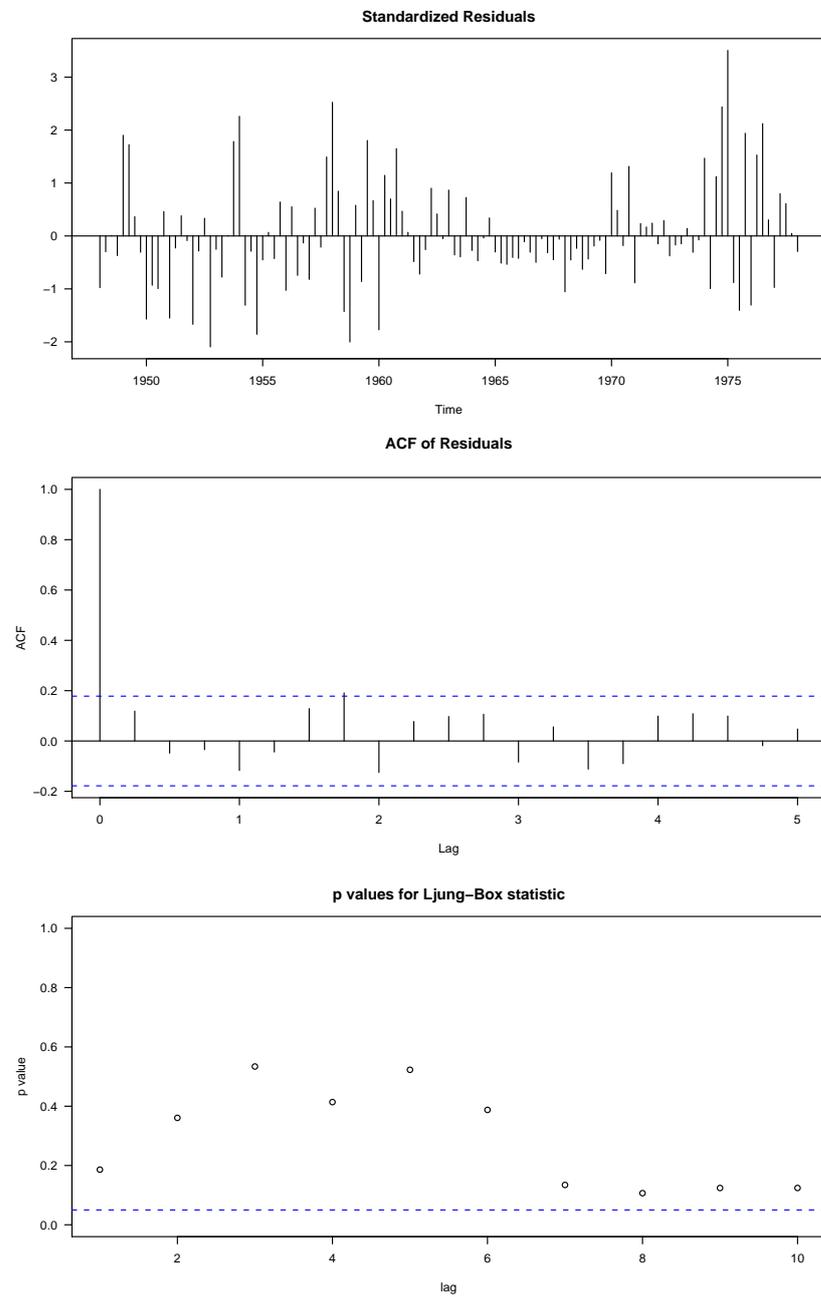


Figure 5.9: Residual diagnostics for the US unemployment series.

```
> z = arima(unemp, order=c(2, 0, 0))
> p = predict(z, n.ahead = 10)
```

The predictions can be viewed with the command

```
> p$pred
      Qtr1      Qtr2      Qtr3      Qtr4
1978          5.812919 5.491268 5.243253
1979 5.067017 4.954379 4.893856 4.872947
1980 4.879708 4.903719 4.936557
```

and their standard errors with

```
> p$se
      Qtr1      Qtr2      Qtr3      Qtr4
1978          0.3572328 0.6589087 0.9095040
1979 1.0969935 1.2248696 1.3040856 1.3479499
1980 1.3689669 1.3771201 1.3792859
```

It can be useful to plot the forecasts on the same graph as the original time series. This a relatively complex task, and would probably be worth packaging up as an R function. Here is how 20 forecasts for the `unemp` series and their standard errors can be plotted

```
> p = predict(z, n.ahead = 20)
> xlim = range(time(unemp), time(p$pred))
> ylim = range(unemp, p$pred - 2 * p$se, p$pred + 2 * p$se)
> plot(unemp, xlim = xlim, ylim = ylim)
> lines(p$pred, lwd=2)
> lines(p$pred - 2 * p$se, lty = "dotted")
> lines(p$pred + 2 * p$se, lty = "dotted")
```

The result of this appears in figure 5.10.

Notice that the standard errors around forecasts widen rapidly and the forecasts themselves seem to be tending to a constant value. If we extend the forecast period, this becomes more obvious as shown in figure 5.11. In fact, over the long term, forecasts for stationary series ultimately converge to the mean of the series and the standard errors for the forecasts tend to the standard deviation of the series. This means that we can only expect to gain advantage from the use of short-term forecasts.

5.5 Seasonal Models

5.5.1 Purely Seasonal Models

Many time series exhibit strong seasonal characteristics. We'll use s to denote the seasonal period. For monthly series, $s = 12$, and for quarterly series $s = 4$. These are the most common cases, but seasonal patterns can show up in other places (e.g. weekly patterns in daily observations or daily patterns in hourly data).

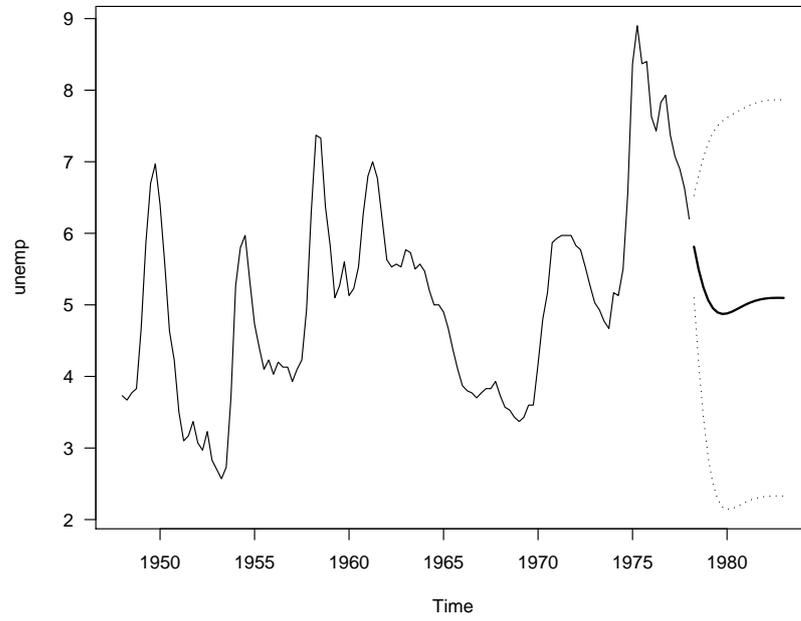


Figure 5.10: Forecasts for the unemployment data.

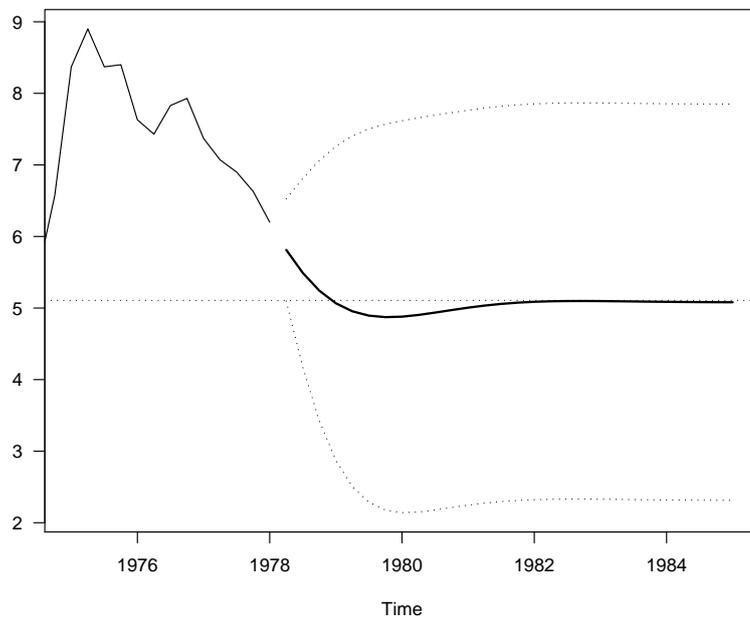


Figure 5.11: Long-term forecasts for the US unemployment series.

Seasonal effects can be modelled by including coefficients at lags which are multiples of the seasonal period. For example, the model

$$Y_t + \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_P Y_{t-Ps} = \varepsilon_t + \Theta_1 \varepsilon_{t-s} + \Theta_2 \varepsilon_{t-2s} + \cdots + \Theta_Q \varepsilon_{t-Qs}$$

is the seasonal analog of an ARMA model. Caution is required with models which only involve coefficients at multiples of the seasonal period because they provide independent models for each of s seasonal sub-series and make no statement about the relationship between the series. For example, the model above applies equally well to the case where the seasonal sub-series are identical

$$Y_{ks+1} = Y_{ks+2} = \cdots = Y_{ks+s}$$

as it does to the case where they are independent.

5.5.2 Models with Short-Term and Seasonal Components

In practise, series will have a mixture of both seasonal and short-term dependencies. It is possible to consider models for such series by including both seasonal and non-seasonal terms. This can be done directly, but it turns out to be more useful to consider a slightly different class of model.

Recall that the general ARMA(p, q) model can be written as

$$\phi(L)Y_t = \theta(L)\varepsilon_t.$$

Each of the $\theta(L)$ and ϕ operators can be written as polynomials in L . These have the factorised form

$$(1 + a_1 L)(1 + a_2 L) \cdots (1 + a_p L)Y_t = (1 + b_1 L)(1 + b_2 L) \cdots (1 + b_q L)\varepsilon_t.$$

Non-stationarity can also be handled by introducing additional factors of the form $(1 - L)$. This suggests a different way of handling seasonal effects. We can simply introduce additional factors of the form $(1 + AL^{12})$ into the autoregressive operator or $(1 + BL^{12})$ into the moving average operator.

The operator on each side of the general model is generally written as a product of a polynomial in L times a polynomial in L^s . The general ARMA model is

$$\Phi(L^s)\phi(L)Y_t = \Theta(L^s)\theta(L)\varepsilon_t,$$

where Φ is a polynomial of degree P , ϕ is a polynomial of degree p , Θ is a polynomial of order Q and θ is a polynomial of degree q . Such a model is denoted as ARMA(p, q) \times (P, Q) $_s$.

Non-stationarity can be accommodated by introducing additional differencing terms. This produces a seasonal ARIMA model, the most general form of which is

$$\Phi(L^s)\phi(L)(1 - L)^d(1 - L^s)^D Y_t = \Theta(L^s)\theta(L)\varepsilon_t,$$

which is denoted by ARMA(p, d, q) \times (P, D, Q) $_s$.

As an example, we'll consider the monthly average temperatures in Dubuque, Iowa, from January 1964 to December 1975. Figure 5.12 shows a plot of the series. It contains a very strong seasonal pattern. The strength of the pattern suggests that we should start by differencing the series at the seasonal lag.

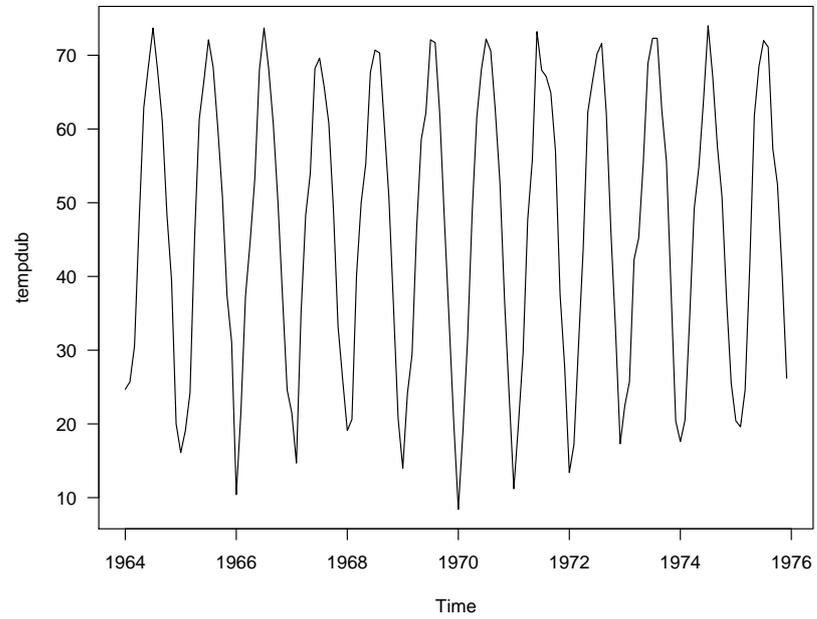


Figure 5.12: Monthly average temperatures °C in Dubuque, Iowa.

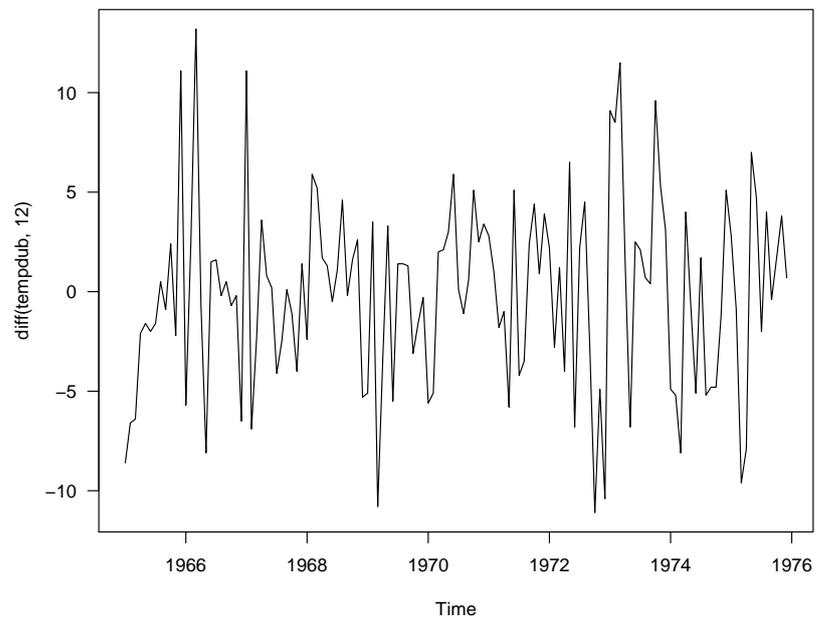


Figure 5.13: The seasonally differenced Dubuque temperature series.

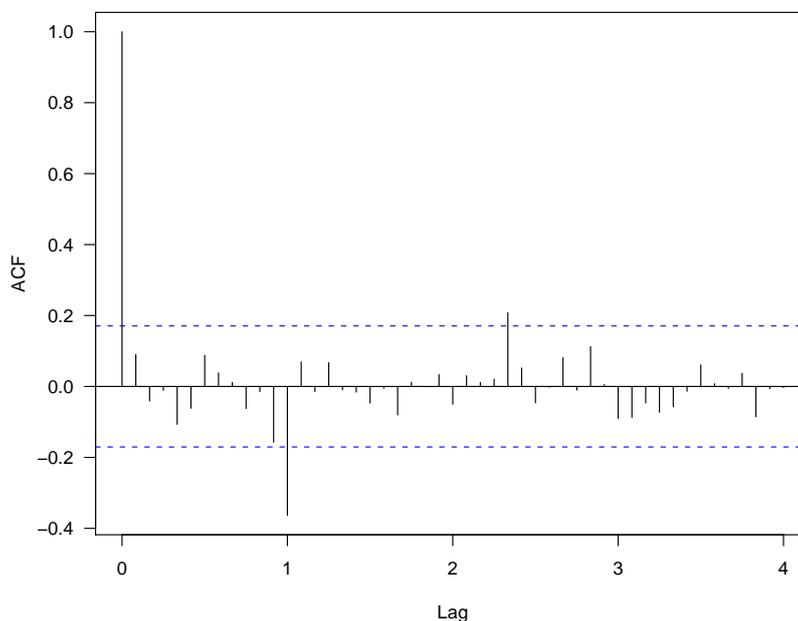


Figure 5.14: The acf of the seasonally differenced Dubuque temperature series.

Figure 5.13 shows a plot of the seasonally differenced series and 5.14 shows its acf function. Together the plots indicate the differenced series is stationary. To decide on an appropriate model we also need to inspect a plot of the pacf of the seasonally differenced series. This is shown in figure 5.15. When taken in conjunction with the structure of the acf, we have clear evidence that a seasonal MA(1) is an appropriate model for the differenced series.

The R function `arima` can be used to fit the full model, which is ARIMA $(0,0,0) \times (0,1,1)_{12}$. Assuming that the series has been read into the variable `tempdub`, the fitting process and display of the results can be carried out as follows.

```
> z = arima(tempdub, seas = c(0,1,1))
> z

Call:
arima(x = tempdub, seasonal = c(0, 1, 1))

Coefficients:
      sma1
    -1.0000
s.e.    0.0967

sigma^2 estimated as 11.69:  log likelihood = -364.48,
      aic = 732.96
```

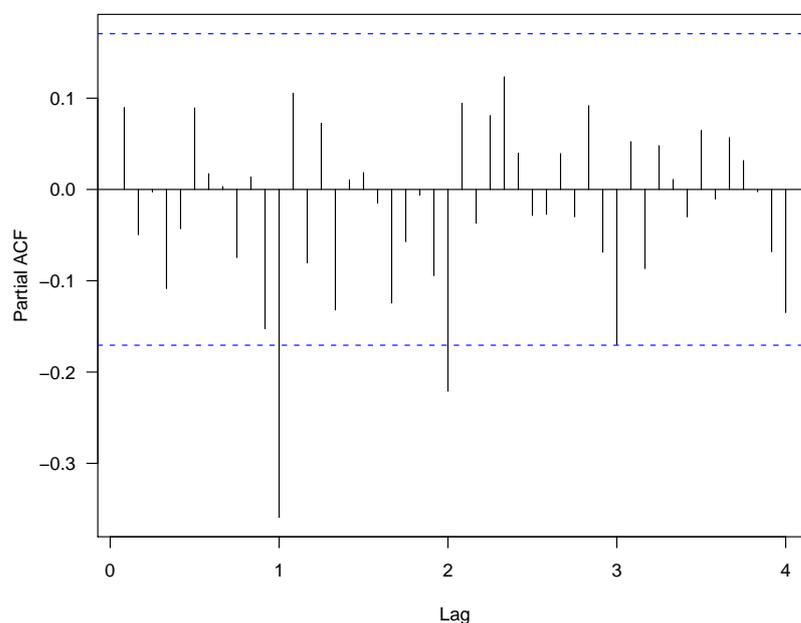


Figure 5.15: The pacf of the seasonally differenced Dubuque temperature series.

Before producing forecasts of the series we need to check the residuals from the fitting process to see that they are (close to) white noise. As before, this can be done with the `tsdiag` function.

```
> tsdiag(z)
```

As shown by figure 5.16, the model seems to fit well.

5.5.3 A More Complex Example

Figure 5.17 shows a plot of a data set presented by Box and Jenkins in their classic time series text. The plot shows the number of international airline passengers, recorded monthly. There is clearly a strong seasonal trend present in the data — the number of passengers peaks during the Northern summer.

There is a clear seasonal effect present in the series, but the size of the seasonal effects seems to be increasing as the level of the series increases. The number of passengers is clearly increasing with time, with the number travelling in July and August always being roughly 50% greater than the number travelling in January and February. This kind of proportional variability suggests that it would be more appropriate to examine the series on a log scale. Figure 5.18 shows the data plotted in this way. On that scale the series shows a consistent level of seasonal variation across time. It seems appropriate to analyse this time series on the log scale.

The transformed series is clearly nonstationary. This means that we need to consider differencing. The series appears to contain both an upward trend and a

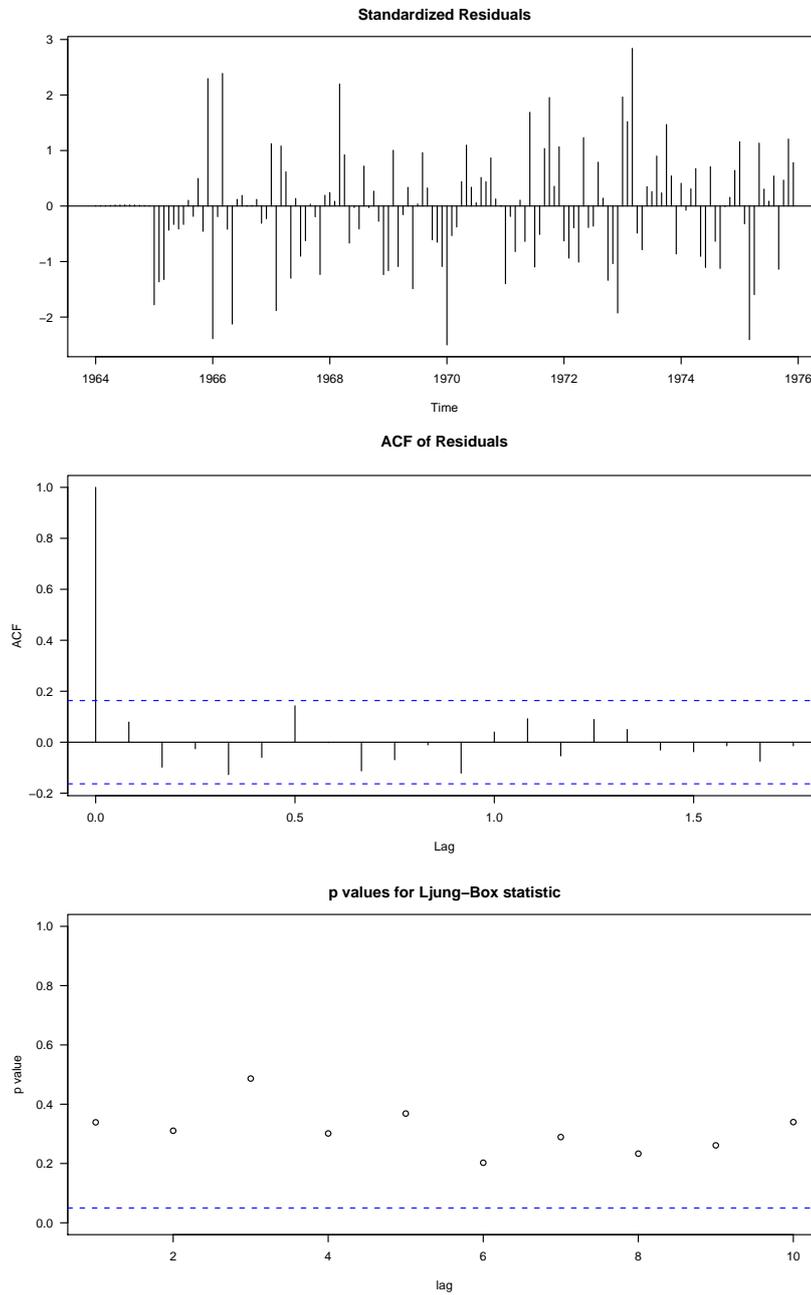


Figure 5.16: Residual diagnostics for the Dubuque temperature series.

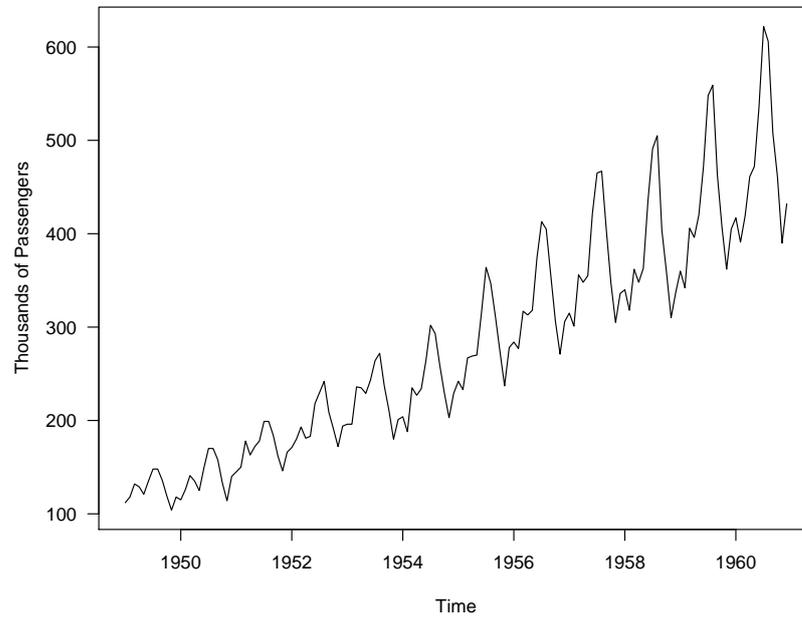


Figure 5.17: International airline passengers, monthly totals (in thousands).

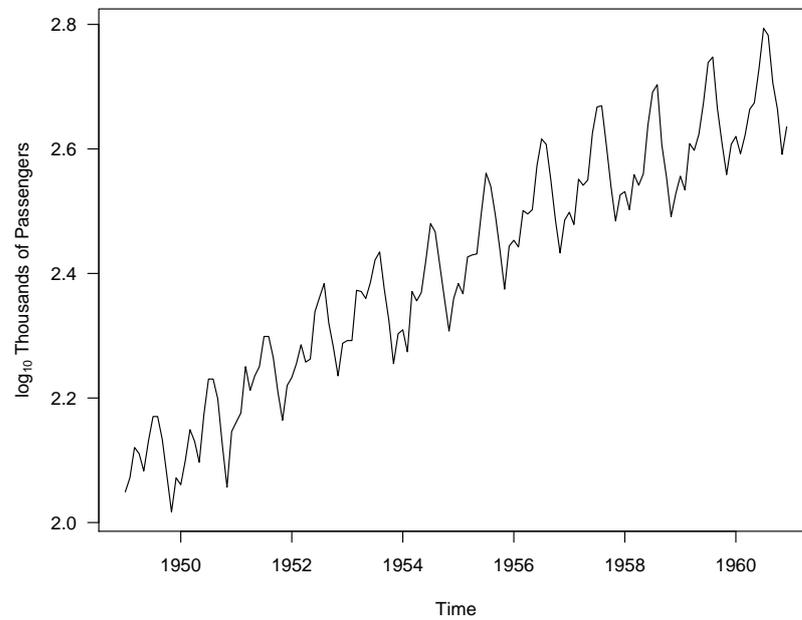


Figure 5.18: The log international airline passenger series.

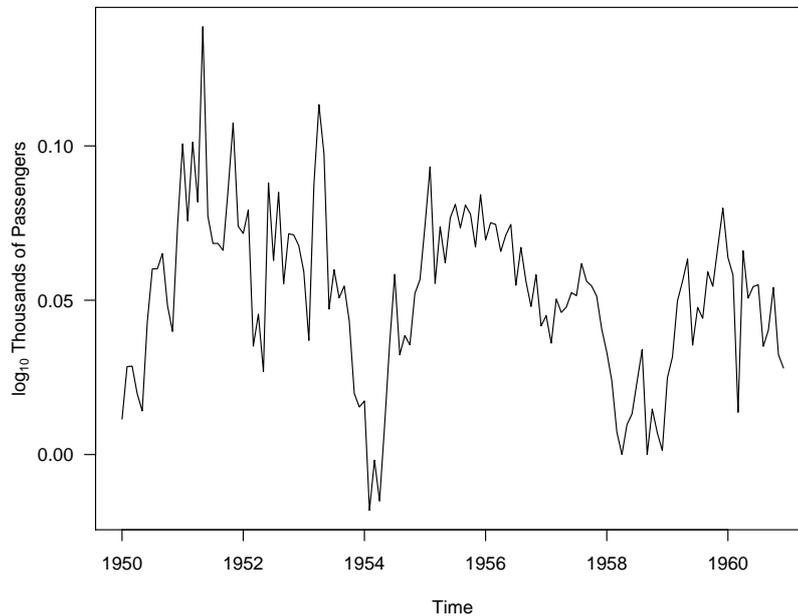


Figure 5.19: The seasonally differenced log international airline passenger series.

seasonal pattern. This means that we need to choose between differencing at lag 1 and at lag 12. If we choose to difference at lag 12 it may be possible to eliminate both trends, so it seems preferable to start with this kind of differencing. The results of the differencing are shown in figure 5.19.

There is a possibility that the differenced series is non-stationary. We need to check this by computing and plotting its acf. This is shown in figure 5.20.

The acf does not appear to be dying out, which suggests that an additional amount of differencing is required. Again, there is a choice between differencing at lag 1 or at lag 12. Because there is no strong seasonal pattern we will try differencing at lag 1. The resulting series and its acf are shown in figures 5.21 and 5.22.

The series now seems to be stationary and so we need to decide on an appropriate model to fit. To do this we need to examine the pacf of the twice differenced series. A plot of this pacf is shown in figure 5.23.

The message conveyed by the acf and pacf functions is not as clear as it has been in previous examples. At least one of the plots must show a mix of decay and oscillation. The “better” candidate for this is the pacf and we will proceed by assuming that this is indeed the case.

There are two large correlations in the acf; at lags of 1 and 12 months. This suggests that an appropriate model for the log passengers series would be $ARIMA(0, 1, 1) \times (0, 1, 1)$. This model can be fitted as follows:

```
> z = arima(log10(airpass), order = c(0,1,1), seas = c(0,1,1))
> z
```

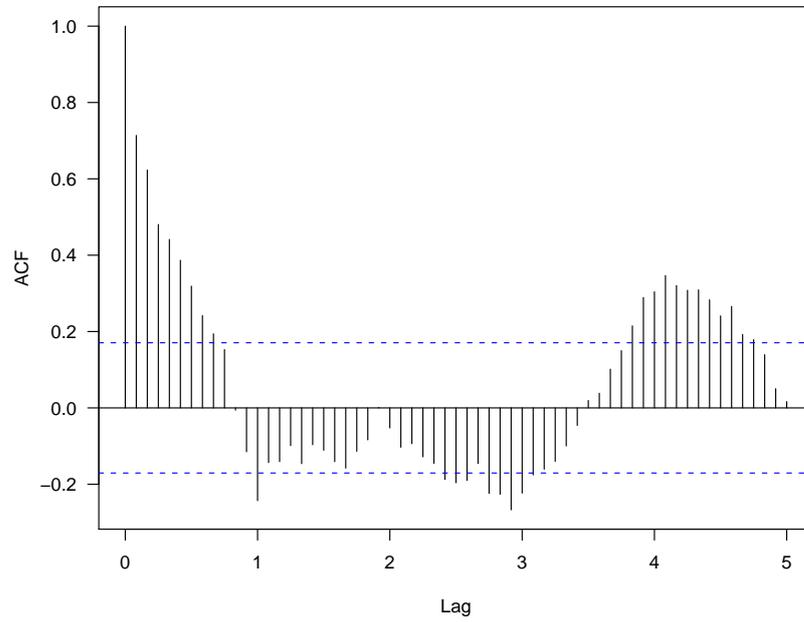


Figure 5.20: The acf of the seasonally differenced log international airline passengers series.

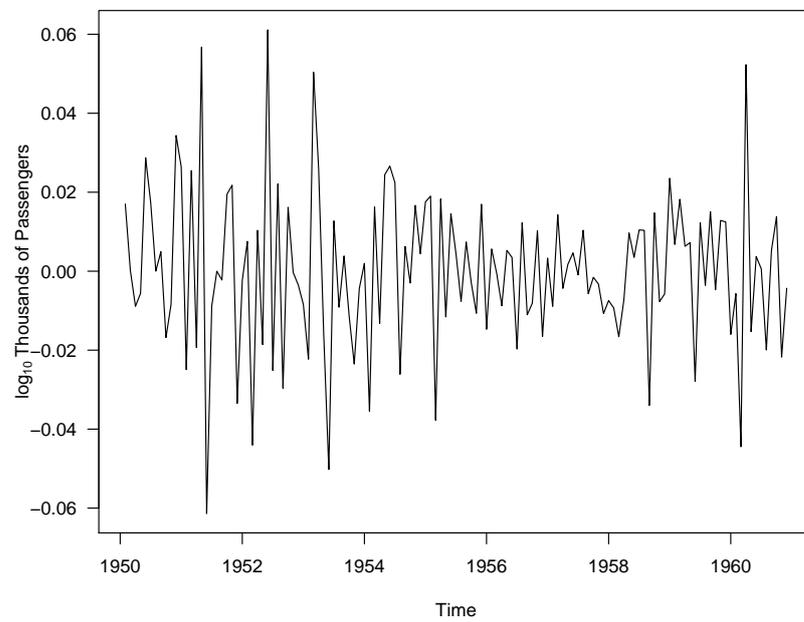


Figure 5.21: The twice differenced log international airline passenger series.

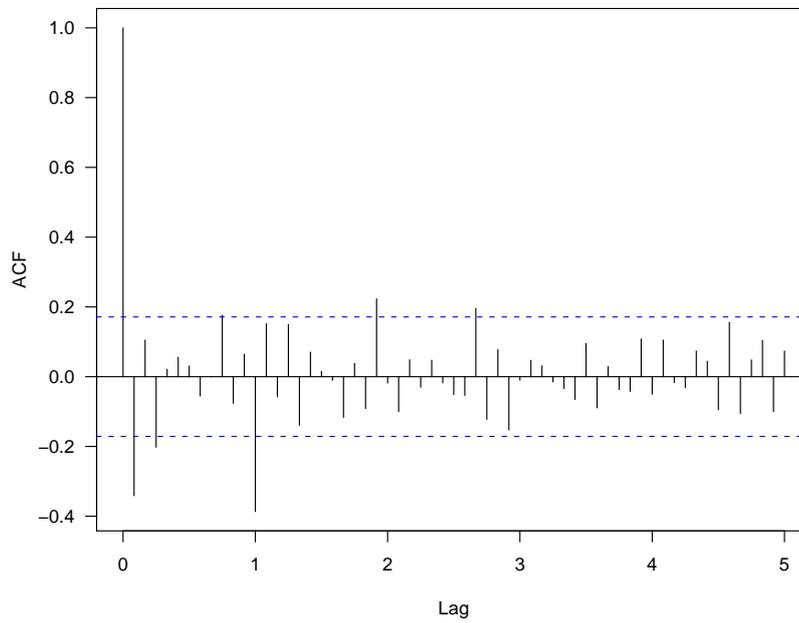


Figure 5.22: The acf of the twice differenced log international airline passengers series.

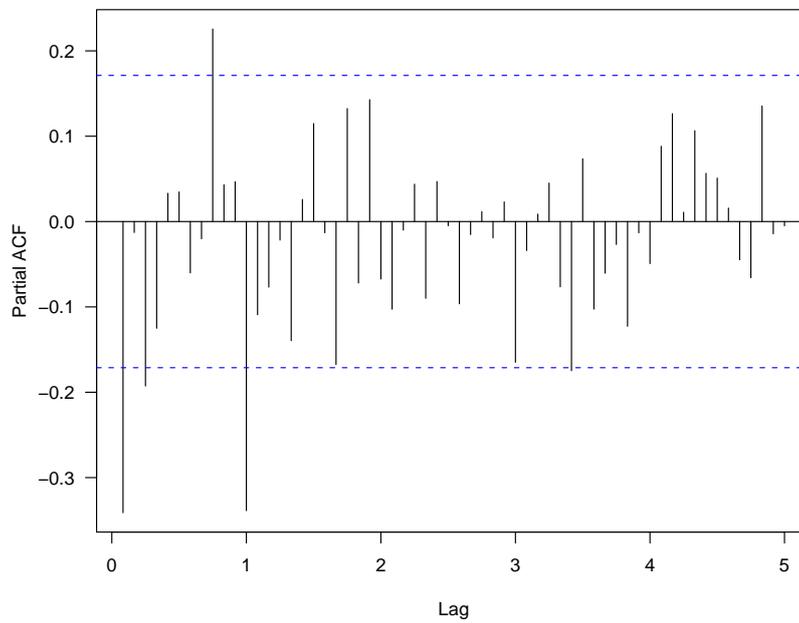


Figure 5.23: The pacf of the twice differenced log international airline passengers series.

```
Call:
arima(x = log10(airpass), order = c(0, 1, 1),
      seasonal = c(0, 1, 1))

Coefficients:
          ma1      sma1
      -0.4018  -0.5569
s.e.    0.0896   0.0731

sigma^2 estimated as 0.0002543:  log likelihood = 353.96,
aic = -701.92
```

Both coefficients are highly significant, so we can move onto examining the residuals with the `tsdiag` function. The results are shown in figure 5.24 and indicate that the model fits well.

Finally, we can obtain and plot forecasts for the series. These are shown in figure 5.25.

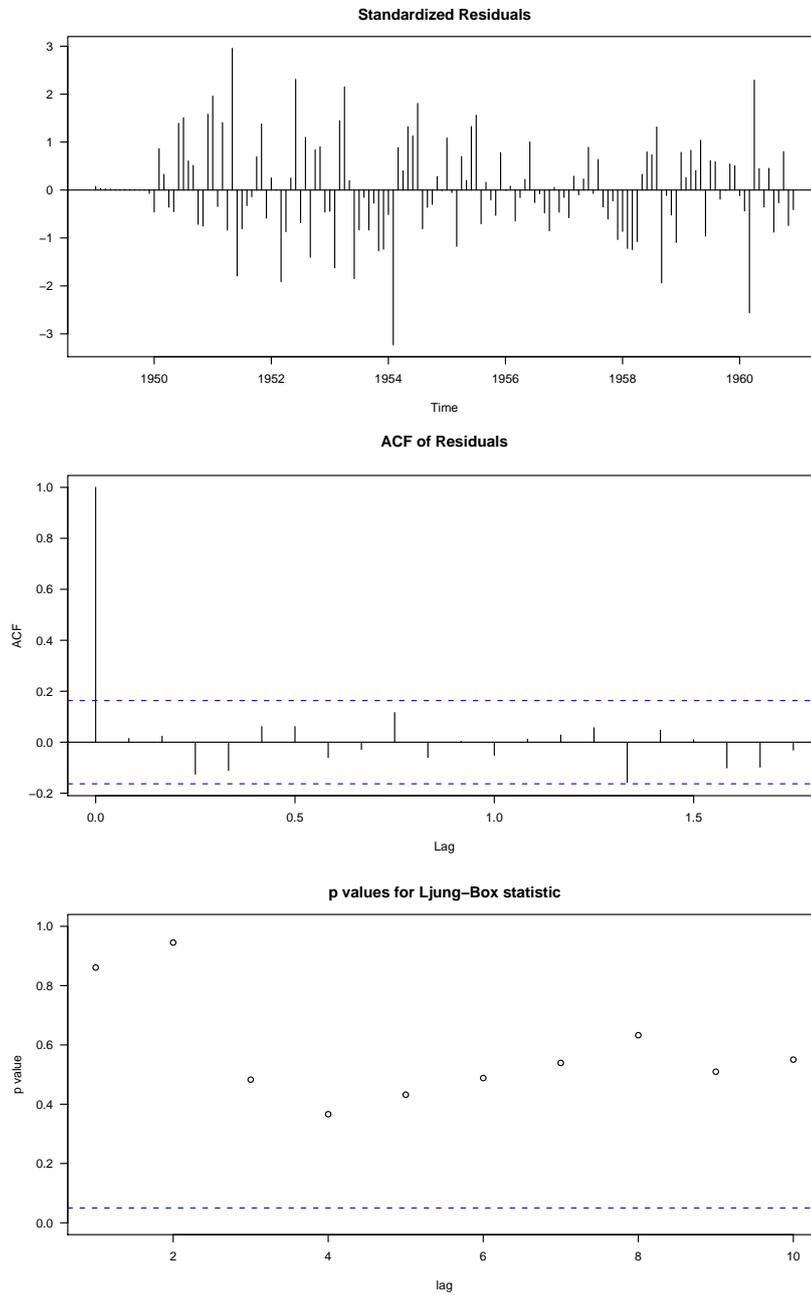


Figure 5.24: Residual diagnostics for the log air passenger series.

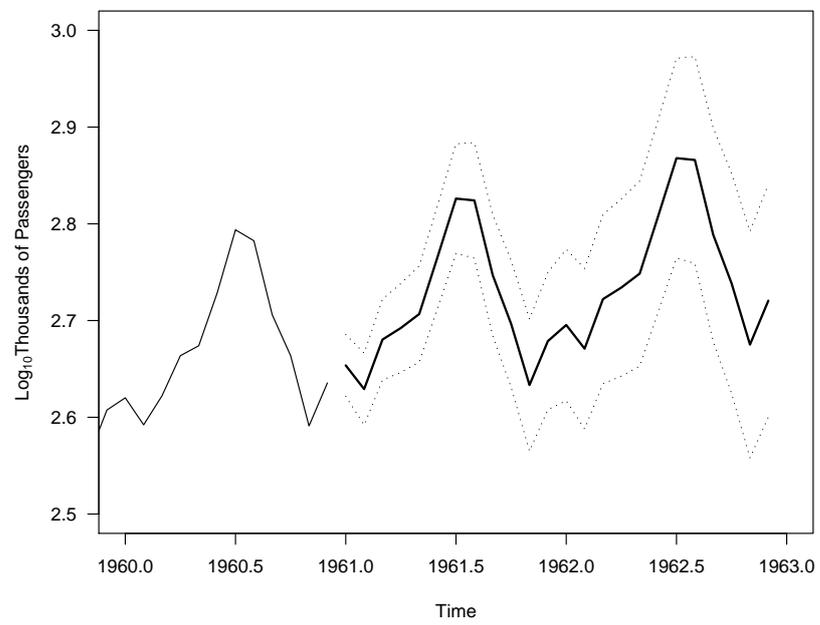


Figure 5.25: Forecasts for the log air passenger series.

Chapter 6

Frequency Domain Analysis

6.1 Some Background

6.1.1 Complex Exponentials, Sines and Cosines

The following formula defines the relationship between the complex exponential function and the real sine and cosine functions.

$$e^{i\theta} = \cos \theta + i \sin \theta$$

From this it is possible to derive many trigonometric identities. For example, we know that

$$e^{i(\theta+\phi)} = \cos(\theta + \phi) + i \sin(\theta + \phi)$$

and also that

$$\begin{aligned} e^{i(\theta+\phi)} &= e^{i\theta} e^{i\phi} \\ &= (\cos \theta + i \sin \theta)(\cos \phi + i \sin \phi) \\ &= \cos \theta \cos \phi - \sin \theta \sin \phi + i(\cos \theta \sin \phi + \sin \theta \cos \phi). \end{aligned}$$

Equating real and imaginary parts

$$\begin{aligned} \cos(\theta + \phi) &= \cos \theta \cos \phi - \sin \theta \sin \phi \\ \sin(\theta + \phi) &= \cos \theta \sin \phi + \sin \theta \cos \phi. \end{aligned}$$

It is also possible to “invert” the basic formula to obtain the following representations for the sine and cosine functions.

$$\begin{aligned} \cos \theta &= \frac{e^{i\theta} + e^{-i\theta}}{2} \\ \sin \theta &= \frac{e^{i\theta} - e^{-i\theta}}{2i}. \end{aligned}$$

For many people it is a natural instinct to try to rewrite $e^{i\theta}$ in the $\cos \theta + i \sin \theta$ form. This is often a mistake because the exponential function is usually easier to handle.

Example 6.1.1 (From Homework)

$$\sum_{t=-T}^T e^{i\lambda t} = \frac{\sin \lambda(T + 1/2)}{\sin \lambda/2}$$

While it is possible to show this by converting immediately to cosines and sines, it is much simpler to recognise that this is just a geometric series and use the formula for summing a geometric series.

$$\begin{aligned} \sum_{t=-T}^T e^{i\lambda t} &= e^{-i\lambda T} \sum_{t=0}^{2T} e^{i\lambda t} \\ &= e^{-i\lambda T} \left(\frac{e^{i\lambda(2T+1)} - 1}{e^{i\lambda} - 1} \right) \\ &= \frac{e^{i\lambda(T+1)} - e^{-i\lambda T}}{e^{i\lambda} - 1} \\ &= \frac{e^{i\lambda(T+1)} - e^{-i\lambda T}}{e^{i\lambda} - 1} \times \frac{e^{-i\lambda/2}}{e^{-i\lambda/2}} \\ &= \frac{e^{i\lambda(T+1/2)} - e^{-i\lambda(T+1/2)}}{e^{i\lambda/2} - e^{-i\lambda/2}} \\ &= \frac{\sin \lambda(T + 1/2)}{\sin \lambda/2} \end{aligned}$$

6.1.2 Properties of Cosinusoids

The general cosinusoid function is defined to be

$$f(t) = a \cos(\lambda t + \phi)$$

where a is the *amplitude*, λ is the *angular frequency* and ϕ is the *phase* of the cosinusoid.

When t takes integer values, it makes sense to restrict λ to the range $[0, 2\pi]$, because

$$\begin{aligned} a \cos((\lambda + 2k\pi)t + \phi) &= a \cos(\lambda t + \phi + 2k\pi t) \\ &= a \cos(\lambda t + \phi) \end{aligned}$$

(\cos is periodic with period 2π and $2k\pi t$ is a multiple of 2π .) This lack of identifiability is known as the *aliasing* problem. It is illustrated in figure 6.1.

Note that

$$a \sin(\lambda t + \phi) = a \cos(\lambda t + (\phi + \pi))$$

so that sines are also cosinusoids. It is also common to refer the function

$$ae^{i(\lambda t + \phi)}$$

as a complex cosinusoid.

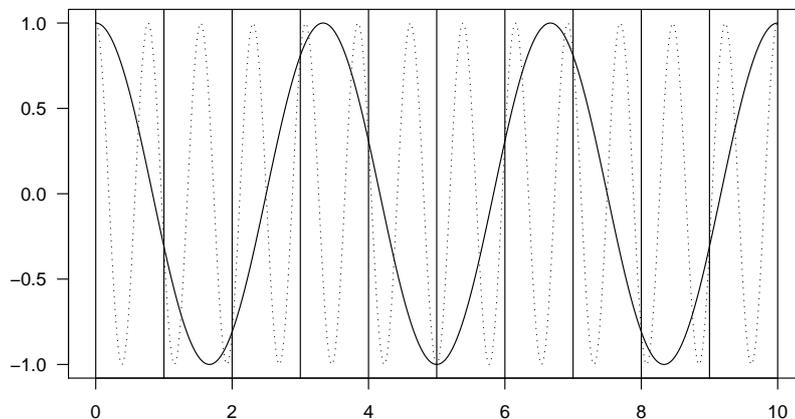


Figure 6.1: The aliasing problem for cosines

6.1.3 Frequency and Angular Frequency

A cosinusoid with frequency π

$$a \cos(\pi t + \phi)$$

repeats itself every two time units. Such a function is usually said to have a frequency of 0.5 because it goes through 0.5 cycles in a unit time period.

In general

$$\text{Frequency} = \frac{\text{Angular Frequency}}{2\pi}.$$

Frequency is more “meaningful” but leads to lots of 2π s in formulae. Because of this, it is usual to carry out theoretical studies of time series with angular frequency, but to perform data analysis in terms of frequency.

Theory	Practise
$a \cos(\lambda t + \phi)$	$a \cos(2\pi\lambda t + \phi)$

The “curse” of time series analysis is that $2\pi \neq 1$.

6.1.4 Invariance and Complex Exponentials

We will be considering “experimental” situations which are in some sense “time-invariant.” This means that we can expect to obtain the same types of results today as we might tomorrow. One definition of invariance for a function f is that it satisfy

$$f(t + u) = f(t) \quad t, u = 0, \pm 1, \pm 2, \dots$$

The only functions which satisfy this condition are constant.

A less restrictive condition would require

$$f(t + u) = C_u f(t) \quad t, u = 0, \pm 1, \pm 2, \dots, \text{ and } C_1 \neq 0 \quad (6.1)$$

For positive t this means

$$f(t) = C_1 f(t - 1) = C_1^2 f(t - 2) = \dots = C_1^t f(0) \quad t \geq 0,$$

while for negative t

$$f(t) = C_1 f(t+1) = C_1^2 f(t+2) = \dots = C_1^t f(0) \quad t \leq 0.$$

This means that for any $t \in \mathbb{Z}$ we can write

$$f(t) = C_1^t f(0).$$

If we write $C_1 = e^\alpha$ (for α real or complex) and $A = f(0)$, then the general solution of equation 6.1 is

$$f(t) = A e^{\alpha t}.$$

The bounded solutions correspond to $\alpha = i\lambda$ for λ real. In other words, the general bounded solution to 6.1 is

$$f(t) = A e^{i\lambda t}.$$

This type of invariance also extends by linearity to functions of the form

$$f(t) = \sum_j c_j e^{i\lambda_j t}$$

because

$$\begin{aligned} f(t+u) &= \sum_j c_j e^{i\lambda_j(t+u)} \\ &= \sum_j c_j e^{i\lambda_j t} e^{i\lambda_j u} \\ &= \sum_j c'_j e^{i\lambda_j t} \end{aligned}$$

The study of functions which can be written as a sum of complex exponentials is called *harmonic analysis* or *Fourier analysis*.

6.2 Filters and Filtering

6.2.1 Filters

A *filter* is an operation which takes one time series as input and produces another as output. We indicate that $Y(t)$ is a filtered version of $X(t)$ as follows.

$$Y(t) = \mathcal{A}[X](t).$$

A filter is called *linear* if it satisfies

$$\mathcal{A}[\alpha X + \beta Y](t) = \alpha \mathcal{A}[X](t) + \beta \mathcal{A}[Y](t)$$

for all constants α and β , and *time-invariant* if it satisfies

$$\mathcal{A}[L^u X](t) = L^u \mathcal{A}[X](t)$$

for all integer values of u .

An important class of linear, time-invariant filters can be written in the form

$$\mathcal{A}[X](t) = \sum_{u=-\infty}^{\infty} a(u) X(t-u).$$

Here the $a(u)$ are called the *filter coefficients*.

Example 6.2.1 Moving Average Filters

Moving average filters of the form

$$\mathcal{A}[X](t) = \frac{1}{2M+1} \sum_{u=-M}^M X(t-u)$$

are used for smoothing time series.

Example 6.2.2 The Differencing Filter

The differencing filter defined by

$$\mathcal{A}[X](t) = X(t) - X(t-1)$$

is used to eliminate long-term trends from time series.

6.2.2 Transfer Functions

While the filter coefficients provide a complete description of what a filter does, they may not provide much intuition about what kind of effect a filter will produce. An alternative way of investigating what a filter does, is to examine its effect on complex exponentials (or equivalently on sine and cosine functions). For notational convenience we will define the function $E^\lambda(t)$ by

$$E^\lambda(t) = e^{i\lambda t}.$$

Clearly,

$$\begin{aligned} L^u E^\lambda(t) &= e^{i\lambda(t+u)} \\ &= e^{i\lambda u} E^\lambda(t). \end{aligned}$$

Time invariance and linearity then allow us to show

$$\begin{aligned} \mathcal{A}[E^\lambda](t+u) &= L^u \mathcal{A}[E^\lambda](t) \\ &= \mathcal{A}[L^u E^\lambda](t) \\ &= \mathcal{A}[e^{i\lambda u} E^\lambda] \\ &= e^{i\lambda u} \mathcal{A}[E^\lambda](t). \end{aligned}$$

Setting $t = 0$ produces

$$\mathcal{A}[E^\lambda](u) = e^{i\lambda u} \mathcal{A}[E^\lambda](0).$$

The function $A(\lambda) = \mathcal{A}[E^\lambda](0)$ is known as the *transfer function* of the filter. The argument above has shown that

$$\mathcal{A}[E^\lambda](u) = A(\lambda) E^\lambda(u).$$

In other words, linear time-invariant filtering of a complex exponential function produces a constant multiple of a complex exponential function with the same frequency.

Note that for any integer value of k ,

$$E^{\lambda+2\pi k}(t) = E^{\lambda}(t)$$

for all $t \in \mathbb{Z}$. This means that transfer functions are periodic with period 2π .

In general, transfer functions are complex-valued. It can often be useful to rewrite them in their polar form

$$A(\lambda) = G(\lambda)e^{i\phi(\lambda)}.$$

$G(\lambda)$ is the *gain function* of the filter and $\phi(\lambda)$ is the *phase function*.

If the filter's coefficients are real-valued then it is easy to show that the transfer function satisfies

$$A(-\lambda) = \overline{A(\lambda)}.$$

This in turn means that the gain must satisfy

$$G(-\lambda) = G(\lambda)$$

and the phase must satisfy

$$\phi(-\lambda) = -\phi(\lambda).$$

6.2.3 Filtering Sines and Cosines

The transfer function describes the effect of a filter on complex exponential functions. Using linearity and the representations of sin and cos in terms of complex exponentials it is easy to show that filtering

$$R \cos(\lambda t + \theta)$$

produces the result

$$G(\lambda)R \cos(\lambda t + \theta + \phi(\lambda)).$$

The gain and phase functions (or equivalently the transfer function) of a filter describe the action of the filter on sinusoids in exactly the same way as they do for complex exponentials.

6.2.4 Filtering General Series

We have seen that a filter's transfer function can be used to describe the effect of the filter on a single complex sinusoid. It can also be used to describe the effect of the filter on more general series.

If a series can be represented in the form

$$X(t) = \sum_j c_j e^{i\lambda_j t},$$

linearity and time invariance mean

$$\mathcal{A}[X](t) = \sum_j A(\lambda_j) c_j e^{i\lambda_j t}.$$

If $A(\lambda_j)$ is small, the component at that frequency will be damped down. If $A(\lambda_j)$ is large, the component will be preserved or amplified.

6.2.5 Computing Transfer Functions

Because transfer functions are a good way of describing the effect of filters, it is useful to have a simple way of computing them. Suppose that the filter coefficients satisfy

$$\sum_{u=-\infty}^{\infty} |a(u)| < \infty$$

then the transfer function is given by

$$A(\lambda) = \sum_{u=-\infty}^{\infty} a(u)e^{-i\lambda u}.$$

We can see this as follows:

$$\begin{aligned} \mathcal{A}[E^\lambda](t) &= \sum_{u=-\infty}^{\infty} a(u)E^\lambda(t-u) \\ &= \sum_{u=-\infty}^{\infty} a(u)e^{i\lambda(t-u)} \\ &= \sum_{u=-\infty}^{\infty} a(u)e^{-i\lambda u}e^{i\lambda t} \\ &= A(\lambda)E^\lambda(t). \end{aligned}$$

In mathematical terms, $A(\lambda)$ is the discrete Fourier transform of the filter coefficients. The summability condition ensures the existence of the transform.

6.2.6 Sequential Filtering

Filters are often applied sequentially when processing time series data and it is often important to understand the combined effect of a sequence of filters. Suppose that \mathcal{A} and \mathcal{B} are filters defined by

$$\begin{aligned} \mathcal{A}[X](t) &= \sum_{u=-\infty}^{\infty} a(u)X(t-u), \\ \mathcal{B}[X](t) &= \sum_{u=-\infty}^{\infty} b(u)X(t-u). \end{aligned}$$

A little algebra shows that the combined effect of \mathcal{B} followed by \mathcal{A} is

$$\mathcal{C}[X](t) = \sum_{u=-\infty}^{\infty} c(u)X(t-u),$$

where

$$c(u) = \sum_{v=-\infty}^{\infty} a(v)b(u-v).$$

The sequence $\{c(u)\}$ is said to be formed by the *convolution* of the sequences $\{a(u)\}$ and $\{b(u)\}$.

Convolutions are comparatively complex and it is difficult to determine what the effect of combining filters might be by just inspecting the coefficient sequence of the convolution. Transfer functions make the task much easier.

$$\begin{aligned}
C(\lambda) &= \sum_{u=-\infty}^{\infty} c(u)e^{-i\lambda u} \\
&= \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} a(v)b(u-v)e^{-i\lambda u} \\
&= \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} a(v)e^{-i\lambda v}b(u-v)e^{-i\lambda(u-v)} \\
&= \sum_{u'=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} a(v)e^{-i\lambda v}b(u')e^{-i\lambda u'} \\
&= \left(\sum_{v=-\infty}^{\infty} a(v)e^{-i\lambda v} \right) \left(\sum_{u'=-\infty}^{\infty} b(u')e^{-i\lambda u'} \right) \\
&= A(\lambda)B(\lambda)
\end{aligned}$$

So, when filters are applied sequentially, their transfer functions multiply. This simple description of what happens when filters are applied sequentially is another reason that transfer functions are important.

6.3 Spectral Theory

6.3.1 The Power Spectrum

Suppose that $X(t)$ is a stationary time series with autocovariance function $\gamma(u)$. If $\gamma(u)$ satisfies

$$\sum_{u=-\infty}^{\infty} |\gamma(u)| < \infty$$

then we define the *power spectrum* of $X(t)$ to be

$$f_{XX}(\lambda) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} \gamma(u)e^{-i\lambda u} \quad (6.2)$$

Because $\gamma(u) = \gamma(-u)$, $f_{XX}(\lambda)$ must be real valued. It is also possible to show that $f_{XX}(\lambda) \geq 0$.

Equation 6.2 can be inverted as follows:

$$\begin{aligned}
\int_0^{2\pi} e^{i\lambda t} f_{XX}(\lambda) d\lambda &= \int_0^{2\pi} e^{i\lambda t} \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} \gamma(u)e^{-i\lambda u} \\
&= \sum_{u=-\infty}^{\infty} \gamma(u) \frac{1}{2\pi} \int_0^{2\pi} e^{i\lambda(t-u)} d\lambda.
\end{aligned}$$

Now

$$\frac{1}{2\pi} \int_0^{2\pi} e^{i\lambda(t-u)} d\lambda = \begin{cases} 1 & \text{if } u = t, \\ 0 & \text{otherwise,} \end{cases}$$

which means that

$$\gamma(u) = \int_0^{2\pi} e^{i\lambda u} f_{XX}(\lambda) d\lambda.$$

This equation is called the *spectral decomposition of the autocovariance function*.

Note that since $\gamma(0) = \text{var}[X(t)]$, the spectral decomposition says that

$$\text{var}[X(t)] = \int_0^{2\pi} f_{XX}(\lambda) d\lambda$$

This equation shows the variability in $X(t)$ being broken down by frequency. To understand the significance of this, we need to see that the time series can be decomposed into independent frequency components.

6.3.2 The Cramér Representation

To define the Cramér representation we have to introduce the idea of *Stieltjes Integration*. The Stieltjes integral of $g(x)$ with respect to $F(x)$ over the interval $[a, b]$, can be thought of as the limit of approximating sums of the form

$$\sum_{n=0}^N g(x_i)[F(x_{i+1}) - F(x_i)],$$

where $a = x_0 < x_1 < \dots < x_N = b$. The notation

$$\int_a^b \phi(x) dF(x)$$

is used to indicate the limiting value.

When $F(x)$ is differentiable with $f(x) = F'(x)$, the definition of derivative means

$$F(x_{i+1}) - F(x_i) \approx f(x_i)(x_{i+1} - x_i)$$

so that the approximating sums have the form

$$\sum_{n=0}^N g(x_i) f(x_i)(x_{i+1} - x_i).$$

In this case

$$\int_a^b \phi(x) dF(x) = \int_a^b \phi(x) f(x) dx.$$

On the other hand, if $F(x)$ is a step function with a step of height c_i at the value u_i , the Stieltjes integral reduces to the sum

$$\sum_i c_i \phi(u_i).$$

The Stieltjes integral has the benefit of unifying both summation and integration.

To state the Cramér representation we need to extend the Stieltjes integral to handle integration against stochastic processes.

Definition 6.3.1 A stochastic process $Z(\lambda)$ on the interval $[0, 2\pi]$ is an indexed set of random variables, which assigns a random variable $Z(\lambda)$ to each $\lambda \in [0, 2\pi]$. The process is said to have uncorrelated (resp. independent) increments if for $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4$, the increments $Z(\lambda_2) - Z(\lambda_1)$ and $Z(\lambda_4) - Z(\lambda_3)$ are uncorrelated (resp. independent).

It is possible to define an integral against such a process as follows:

$$\int_0^{2\pi} \phi(\lambda) dZ(\lambda) = \text{l.i.m.} \sum_{n=0}^{N-1} \phi\left(\frac{2\pi n}{N}\right) \left[Z\left(\frac{2\pi(n+1)}{N}\right) - Z\left(\frac{2\pi n}{N}\right) \right].$$

The Cramér representation says that it is possible to represent a stationary time series $X(t)$ in the form

$$X(t) = \int_0^{2\pi} e^{i\lambda t} dZ_X(\lambda)$$

for a stochastic process $Z_X(\lambda)$ with uncorrelated increments.

The process $Z_X(\lambda)$ is defined as follows. First we define

$$d_X^T(\lambda) = \sum_{u=-T}^T X(u) e^{-i\lambda u}$$

and then

$$Z_X^T(\lambda) = \int_0^\lambda d_X^T(\alpha) d\alpha = \sum_{u=-T}^T X(u) \left(\frac{1 - e^{-i\lambda u}}{-iu} \right).$$

Finally we let $T \rightarrow \infty$ and set

$$Z_X(\lambda) = \lim_{T \rightarrow \infty} Z_X^T(\lambda).$$

It is possible to show that $Z_X(\lambda)$ is a well-defined stochastic process on $[0, 2\pi]$ which has uncorrelated increments and satisfies symmetry properties such as

$$\begin{aligned} Z_X(\lambda + 2\pi) &= Z_X(\lambda) \\ Z_X(-\lambda) &= \overline{Z_X(\lambda)} \end{aligned}$$

Most important of all, it is possible to make statements about the variability of the increments of $Z_X(\lambda)$. For what we are interested in, it will be necessary to assume that the series $X(t)$ satisfies the additional *mixing* condition

$$\sum_{u=-\infty}^{\infty} |\gamma(u)| < \infty.$$

This says that values of $X(t)$ which are well separated in time will be close to uncorrelated.

If the series is mixing, the variability of the quantities

$$dZ_X(\lambda) = Z_X(\lambda + d\lambda) - Z_X(\lambda).$$

can be written in terms of the power spectrum $f_{XX}(\lambda)$.

$$\text{var}(dZ_X(\lambda)) = f_{XX}(\lambda)(d\lambda)^2$$

This together with the uncorrelatedness of the increments can be written in the operational form

$$\text{cov}(dZ_X(\lambda), dZ_X(\mu)) = \delta(\lambda - \mu) f_{XX}(\lambda) d\lambda d\mu, \quad (6.3)$$

where $\delta(u)$ is the *Dirac delta function*, which is defined so that

$$\int \phi(u) \delta(u - u_0) du = \phi(u_0)$$

Equation 6.3 is understood to mean

$$\begin{aligned} \text{cov}\left(\int_0^{2\pi} \phi(\lambda) dZ_X(\lambda), \int_0^{2\pi} \psi(\lambda) dZ_X(\lambda)\right) &= \int_0^{2\pi} \int_0^{2\pi} \phi(\lambda) \overline{\psi(\mu)} \text{cov}(dZ_X(\lambda), dZ_X(\mu)) \\ &= \int_0^{2\pi} \int_0^{2\pi} \phi(\lambda) \overline{\psi(\mu)} \delta(\lambda - \mu) f_{XX}(\lambda) d\lambda d\mu \\ &= \int_0^{2\pi} \phi(\lambda) \overline{\psi(\lambda)} f_{XX}(\lambda) d\lambda \end{aligned}$$

6.3.3 Using The Cramér Representation

The Cramér representation says that we can approximate the time series $X(t)$ to any level of accuracy by an approximation of the form

$$X(t) \approx \sum_{n=0}^N \exp\left(i \frac{2\pi n}{N} t\right) Z_n$$

where the Z_n are uncorrelated random variables. This provides an intuitive way of handling time series theory.

As an example, let's suppose that the series $X(t)$ has the Cramér representation

$$X(t) = \int_0^{2\pi} e^{i\lambda t} d\lambda$$

and that we filter $X(t)$ with a filter which has transfer function $A(\lambda)$ to obtain

$$Y(t) = \mathcal{A}[X](t).$$

The Cramér representation says that we arbitrarily approximate the series $X(t)$ with the sum

$$\sum_{n=0}^N \exp\left(i \frac{2\pi n}{N} t\right) Z_n,$$

where the Z_n are independent random variables. When this approximating series is filtered, the result is (by linearity)

$$\sum_{n=0}^N A\left(\frac{2\pi n}{N}\right) \exp\left(i \frac{2\pi n}{N} t\right) Z_n.$$

On taking limits, we obtain

$$Y(t) = \int_0^{2\pi} e^{i\lambda t} A(\lambda) dZ_X(\lambda).$$

providing a spectral representation of $Y(t)$.

We can also use the Cramér representation to obtain a formula for the power spectrum of the $Y(t)$ series. The representation above says that

$$dZ_Y(\lambda) = A(\lambda)dZ_X(\lambda),$$

and since

$$\begin{aligned} \text{var}\left(dZ_Y(\lambda)\right) &= \text{cov}\left(dZ_Y(\lambda), dZ_Y(\lambda)\right) \\ &= \text{cov}\left(A(\lambda)dZ_X(\lambda), A(\lambda)dZ_X(\lambda)\right) \\ &= A(\lambda)\overline{A(\lambda)}\text{cov}\left(dZ_X(\lambda), dZ_X(\lambda)\right) \\ &= |A(\lambda)|^2\text{var}\left(dZ_X(\lambda)\right) \end{aligned}$$

we immediately see that the power spectrum for $Y(t)$ must be

$$f_{YY}(\lambda) = |A(\lambda)|^2 f_{XX}(\lambda).$$

6.3.4 Power Spectrum Examples

Because of its relationship with the Cramér representation, the power spectrum is a fundamental parameter of interest when dealing with stationary time series. It can be useful to see how the power spectrum behaves for some of the stationary series we have encountered.

Example 6.3.1 (White Noise)

The autocovariance function of a white-noise series $X(t)$ is given by

$$\gamma(u) = \begin{cases} \sigma^2 & u = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The power spectrum can be computed directly from this.

$$\begin{aligned} f_{XX}(\lambda) &= \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} e^{-i\lambda u} \gamma(u) \\ &= \frac{\sigma^2}{2\pi} \end{aligned}$$

This says that a white-noise series is composed of “equal amounts of every frequency.” The name, white noise, comes from an analogy with white light, which contains equal amounts of every colour (wavelength).

Example 6.3.2 The MA(1) Series.

The MA(1) series is defined by

$$Y(t) = \varepsilon(t) + \theta\varepsilon(t-1)$$

and has an autocovariance function defined by

$$\gamma(u) = \begin{cases} \sigma^2(1 + \theta^2) & u = 0, \\ \sigma^2\theta & u = \pm 1, \\ 0 & \text{otherwise.} \end{cases}$$

The power spectrum is thus

$$\begin{aligned} f_{YY}(\lambda) &= \frac{\sigma^2}{2\pi} (\theta e^{-i\lambda} + (1 + \theta^2) + \theta e^{i\lambda}) \\ &= \frac{\sigma^2}{2\pi} (1 + \theta^2 + 2\theta \cos \lambda) \end{aligned}$$

Example 6.3.3 The AR(1) Series.

The AR(1) series is defined by

$$Y(t) = \phi Y(t-1) + \varepsilon(t)$$

and has an autocovariance function defined by

$$\gamma(u) = \sigma^2 \phi^{|u|}$$

The power spectrum is

$$\begin{aligned} f_{YY}(\lambda) &= \frac{\sigma^2}{2\pi} \sum_{u=-\infty}^{\infty} e^{-i\lambda u} \phi^{|u|} \\ &= \frac{\sigma^2}{2\pi} \frac{1}{1 + \theta^2 - 2\phi \cos \lambda}. \end{aligned}$$

The last equality can be shown by direct algebraic manipulation. Alternatively, we can use the fact that $Y(t)$ can be filtered to obtain white noise.

$$Y(t) - \phi Y(t-1) = \varepsilon(t)$$

If $A(\lambda)$ is the transfer function of the filter with coefficients

$$a(u) = \begin{cases} 1 & u = 0, \\ \phi & u = 1, \\ 0 & \text{otherwise.} \end{cases}$$

then

$$|A(\lambda)|^2 f_{YY}(\lambda) = f_{\varepsilon\varepsilon}(\lambda) = \frac{\sigma^2}{2\pi},$$

or equivalently,

$$f_{YY}(\lambda) = \frac{\sigma^2}{2\pi} \frac{1}{|A(\lambda)|^2}.$$

The transfer function is easily seen to be

$$A(\lambda) = 1 - \phi e^{-i\lambda}$$

and hence

$$\begin{aligned} |A(\lambda)|^2 &= A(\lambda)\overline{A(\lambda)} \\ &= (1 - \phi e^{-i\lambda})(1 - \phi e^{i\lambda}) \\ &= 1 + \phi^2 - 2\phi \cos \lambda \end{aligned}$$

from which the result follows.

6.4 Statistical Inference

6.4.1 Some Distribution Theory

In the frequency domain, inference is based on the discrete Fourier transform

$$d_X^T(\lambda) = \sum_{t=0}^{T-1} X(t)e^{-i\lambda t}$$

of a set of T data values $X(0), \dots, X(T-1)$. The discrete Fourier transform has the following analytic properties:

$$\begin{aligned} d_X^T(0) &= \sum_{t=0}^{T-1} X(t) \\ d_X^T(\lambda + 2\pi) &= d_X^T(\lambda) \\ d_X^T(-\lambda) &= \overline{d_X^T(\lambda)} \\ d_X^T(2\pi - \lambda) &= \overline{d_X^T(-\lambda)} \end{aligned}$$

This means that all the information contained in the discrete Fourier transform can be displayed by plotting $d_X^T(\lambda)$ over the interval $[0, \pi]$.

Most importantly, although the discrete Fourier transform is a complex transformation of the values $X(0), \dots, X(T-1)$, it can be computed rapidly using a *fast Fourier transform* (FFT) algorithm. The FFT computes the values of $d_X^T(\lambda)$ for the discrete set of frequencies

$$\lambda_t = \frac{2\pi t}{T}, \quad t = 0, \dots, T-1.$$

Many statistical software packages contain an implementation of the FFT.

Theorem 6.4.1 (*The Distribution of the Fourier Transform*). If $X(t)$ is a stationary, mean-zero time series which satisfies the mixing condition

$$\sum_{u=-\infty}^{\infty} |\gamma(u)| < \infty$$

and has power spectrum $f_{XX}(\lambda)$, then

$$\begin{aligned} \mathbb{E}\left(d_X^T(\lambda)\right) &= 0 \\ \text{var}\left(\frac{d_X^T(\lambda)}{\sqrt{2\pi T}}\right) &\rightarrow f_{XX}(\lambda) \\ \text{cov}\left(\frac{d_X^T(\lambda)}{\sqrt{2\pi T}}, \frac{d_X^T(\mu)}{\sqrt{2\pi T}}\right) &\rightarrow 0, \quad \text{provided, } \lambda \neq \mu. \end{aligned}$$

Finally, under some additional mixing conditions (involving higher order moments), $d_X^T(\lambda)$ is asymptotically (complex) normal.

A complete proof of this result can be found David Brillinger's book *Time Series: Data Analysis and Theory*. We will just show how some of the proof works.

First, by linearity, it is trivial to show that

$$\mathbb{E}\left[d_X^T(\lambda)\right] = 0.$$

The variance result can be shown as follows

$$\begin{aligned} \mathbb{E}\left|\frac{d_X^T(\lambda)}{\sqrt{2\pi T}}\right|^2 &= \mathbb{E}\left[\frac{d_X^T(\lambda)\overline{d_X^T(\lambda)}}{2\pi T}\right] \\ &= \frac{1}{2\pi T} \sum_{t=0}^{T-1} \sum_{s=0}^{T-1} \mathbb{E}\left[X(t)X(s)\right] e^{-i\lambda t} e^{-i\lambda s} \\ &= \frac{1}{2\pi T} \sum_{t=0}^{T-1} \sum_{s=0}^{T-1} \gamma(t-s) e^{-i\lambda(t-s)} \\ &= \frac{1}{2\pi T} \sum_{u=-T+1}^{T-1} (T-|u|)\gamma(u) e^{-i\lambda u} \\ &= \frac{1}{2\pi} \sum_{u=-T+1}^{T-1} \left(1 - \frac{|u|}{T}\right) \gamma(u) e^{-i\lambda u} \\ &\rightarrow f_{XX}(\lambda) \end{aligned}$$

The convergence follows because because $(1 - |u|/T) \uparrow 1$ as $T \rightarrow \infty$.

To show the covariance result we need the preliminary technical result that if $\nu \neq 0$ then the value of

$$\left| \sum_{s=N_1}^{N_2} e^{-i\nu s} \right|$$

is bounded by some constant B_ν . The result follows because

$$\sum_{s=0}^T e^{-i\nu s} = \frac{1 - e^{-i\nu T}}{1 - e^{-i\nu}}$$

which converges to 0 as $T \rightarrow \infty$ provided $\nu \neq 0$.

$$\begin{aligned} \mathbb{E} \left[\frac{d_X^T(\lambda)}{\sqrt{2\pi T}}, \frac{d_X^T(\mu)}{\sqrt{2\pi T}} \right] &= \frac{1}{2\pi T} \sum_{t=0}^{T-1} \sum_{s=0}^{T-1} \mathbb{E} [X(t)X(s)] e^{-i\lambda t} e^{-i\mu s} \\ &= \frac{1}{2\pi T} \sum_{t=0}^{T-1} \sum_{s=0}^{T-1} \gamma(t-s) e^{-i\lambda t} e^{-i\mu s} \\ &= \frac{1}{2\pi T} \sum_{t=0}^{T-1} \sum_{s=0}^{T-1} \gamma(t-s) e^{-i(\lambda t - \mu s)} \\ &= \frac{1}{2\pi T} \sum_{t=0}^{T-1} \sum_{s=0}^{T-1} \gamma(t-s) e^{-i\lambda(t-s)} e^{-i(\lambda-\mu)s} \\ &= \frac{1}{2\pi T} \sum_{u=-T+1}^{T-1} \sum_{s=\max(0,-u)}^{\min(T-1, T-1-u)} \gamma(u) e^{-i(\lambda-\mu)s} e^{-i\lambda u} \\ &= \frac{1}{2\pi T} \sum_{u=-T+1}^{T-1} \gamma(u) e^{-i\lambda u} \sum_{s=\max(0,-u)}^{\min(T-1, T-1-u)} e^{i(\lambda-\mu)s} \end{aligned}$$

This says that

$$\left| \mathbb{E} \left[\frac{d_X^T(\lambda)}{\sqrt{2\pi T}}, \frac{d_X^T(\mu)}{\sqrt{2\pi T}} \right] \right| \leq \frac{B_\lambda}{T} \left| \frac{1}{2\pi} \sum_{u=-T+1}^{T-1} \gamma(u) e^{-i\lambda u} \right|$$

The sum on the right converges to $f_{XX}(\lambda)$, so the whole right hand side must converge to 0.

If the original time series $X(t)$ is normally distributed, then by linearity, $d_X^T(\lambda)$ must also be normally distributed. The general result can be found in Brillinger's book.

6.4.2 The Periodogram and its Distribution

Asymptotically, the discrete Fourier transform has a complex normal distribution with mean 0 and variance proportional to the power spectrum.

$$d_X^T(\lambda) \sim \text{AN}^c(0, 2\pi T f_{XX}(\lambda)).$$

This says that as $T \rightarrow \infty$,

$$\mathbb{E} \left[\frac{1}{2\pi T} \left| d_X^T(\lambda) \right|^2 \right] \rightarrow f_{XX}(\lambda). \quad (6.4)$$

The quantity

$$I_{XX}^T(\lambda) = \frac{1}{2\pi T} \left| d_X^T(\lambda) \right|^2$$

is called the *periodogram*. Equation 6.4 says that the periodogram provides an asymptotically unbiased estimator of the power spectrum.

It is relatively easy to derive the asymptotic distribution of the periodogram. We know that as $T \rightarrow \infty$,

$$\frac{d_X^T(\lambda)}{\sqrt{2\pi T}} \rightarrow N^c(0, f_{XX}(\lambda)).$$

We can represent this limiting distribution in the form

$$(U + iV)\sqrt{f_{XX}(\lambda)/2}$$

where U and V are independent (real-valued) normals, each with mean equal to 0 and variance equal to 1. The limiting distribution of the periodogram can then be written as

$$(U^2 + V^2) \frac{f_{XX}(\lambda)}{2}$$

Because U and V are standard normals, the distribution of $U^2 + V^2$ is χ_2^2 , or an exponential distribution with mean 2. This means that the asymptotic distribution of $I_{XX}^T(\lambda)$ is exponential with mean $f_{XX}(\lambda)$.

There are two important consequences which follow because of the asymptotic distribution of the periodogram. First, $I_{XX}^T(\lambda)$ is not a consistent estimator of $f_{XX}(\lambda)$. Second, the asymptotic variance of the periodogram is $f_{XX}(\lambda)^2$. This means that from a data analytic standpoint, it is better to draw graphs of $\log I_{XX}^T(\lambda)$ rather than $I_{XX}^T(\lambda)$ itself (log is the variance stabilising transformation for the exponential).

6.4.3 An Example – Sunspot Numbers

During the early 1600s Galileo used his invention, the astronomical telescope, to study the heavens. One of the objects he investigated was the sun, and a focus of his study the phenomenon of sunspots. Sunspots are dark areas which appear from time to time on the surface of the sun. They range in size from about 1,500km to 150,000km across. The phenomenon had been known to earlier Greek and Chinese astronomers, but Galileo was the first to systematically study and write about them.

Sunspots are now believed to be areas where the convection currents rising up from the interior of the sun are inhibited by its strong magnetic fields. The temperatures at the centre of a sunspot can be some 2000°K below the usual 6,000°K surface temperature of the sun.

It is now known that the number of sunspots observed on the sun varies with an irregular 11 year period. An index of the level of sunspot has been kept for hundreds of years. The index is related to both the number of sunspots and the area they cover. A value of 100 is high and anything above 150 is regarded as unusually high. The sunspot series is displayed in figure 6.2. From a statistical point of view, it is desirable to transform the series to make the variability

independent of the mean. Because the index is based on counts, variance stabilisation can be carried out by taking square roots. The transformed series is also shown in figure 6.2.

Because the sun has a direct effect on the earth's climate the sunspot cycle has been the subject of intense study by climatologists and astronomers. Because it provides a direct measure of solar activity, the sunspots series has been studied intensively to see if it can be used to predict medium and long-term climate trends.

The periodogram is a useful tool to use when looking for periodicities. Figures 6.3 and 6.4 show the periodogram computed for the square-root of the sunspot index. The vertical lines in the second of these figures correspond to periodicities of 5.5 years, 11 years and 76 years. The first two of these periodicities correspond directly to the 11 year sunspot cycle. The 5.5 yearly period is a harmonic which is produced because the sunspot index increases more rapidly than it declines. The 76 year periodicity corresponds to a basic oscillation in the size of the sun. The observed diameter of the sun varies by about 140km over a 76 year cycle.

6.4.4 Estimating The Power Spectrum

The periodogram provides an asymptotically unbiased estimator of the power spectrum, but one which is not consistent. However, it is possible to derive consistent estimators from it.

Assume that the power spectrum $f_{XX}(\lambda)$ is near-constant in the region of the frequency λ_0 . This says that values of $I_{XX}(\lambda)$ with λ close to λ_0 will be approximately i.i.d. with mean $f_{XX}(\lambda_0)$. Averaging these values will provide an estimate of $f_{XX}(\lambda_0)$ which may have a small amount of bias, but which will have a smaller variance. By trading off variability for bias, we can potentially obtain a much better estimator.

We proceed by choosing a sequence of positive *bandwidth* values B_T , and taking the average of the periodogram values in the interval from $\lambda_0 - B_T/2$ to $\lambda_0 + B_T/2$. This produces a sequence of spectral estimates $f_{XX}^T(\lambda_0)$. If B_T is chosen so that $B_T \downarrow 0$, but at a rate more slowly than $1/T$ and if the power spectrum is smooth in the vicinity of λ_0 , then the sequence $f_{XX}^T(\lambda_0)$ will converge to $f_{XX}(\lambda_0)$. Such *smoothed periodogram* estimators provide a consistent way of estimating the power spectrum. In practise we can either nominate a value for B_T , or specify the number of values to be averaged.

The distribution of smoothed periodogram estimates has a simple asymptotic form. If k periodogram values are averaged to obtain the estimate then the estimate will have an approximate distribution which is $f_{XX}(\lambda_0)\chi_{2k}^2$ distribution (because it is the sum of independent $f_{XX}(\lambda_0)\chi_2^2$ variables).

This distribution has a variance which is proportional to its mean and so such estimates are better viewed on a logarithmic scale. Taking logs also has the advantage that there is a single standard error which applies across all frequencies.

To compute an estimate of the spectrum we must first specify the amount of smoothing to be used, either with a bandwidth or a specification of the number of values to be averaged. There is no firm rule for choosing the amount of smoothing and it is best to try a number of values. In regions where the spectrum

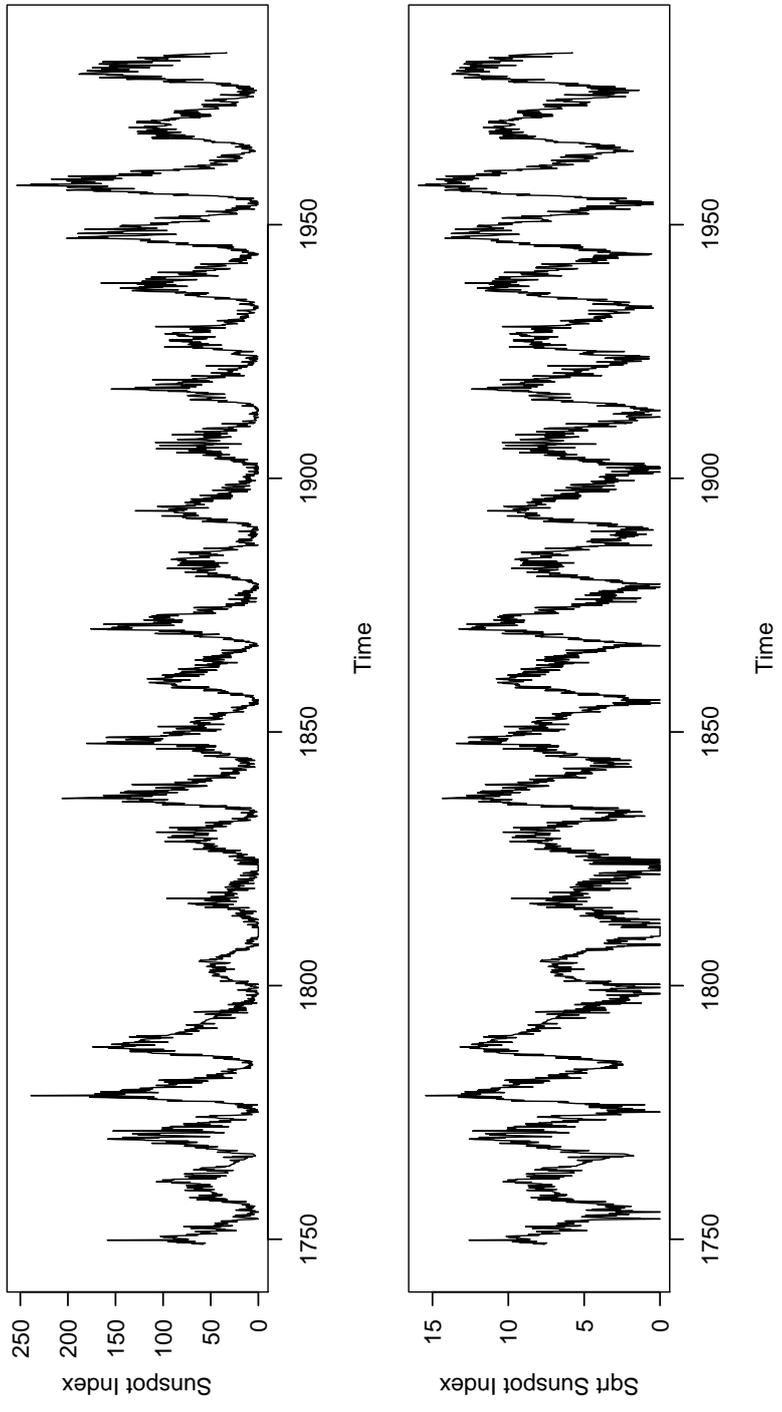


Figure 6.2: The sunspot index series and its square root.

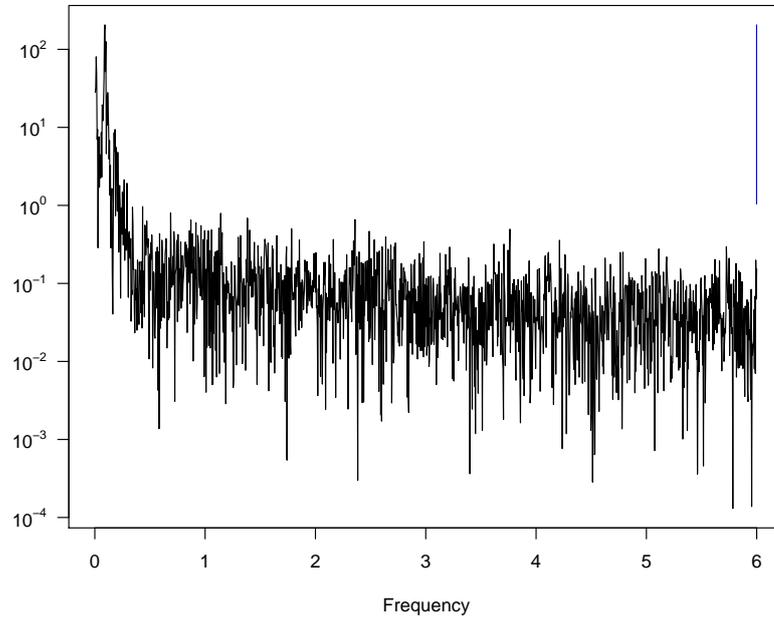


Figure 6.3: The log periodogram computed for the square-root of the sunspot index series.

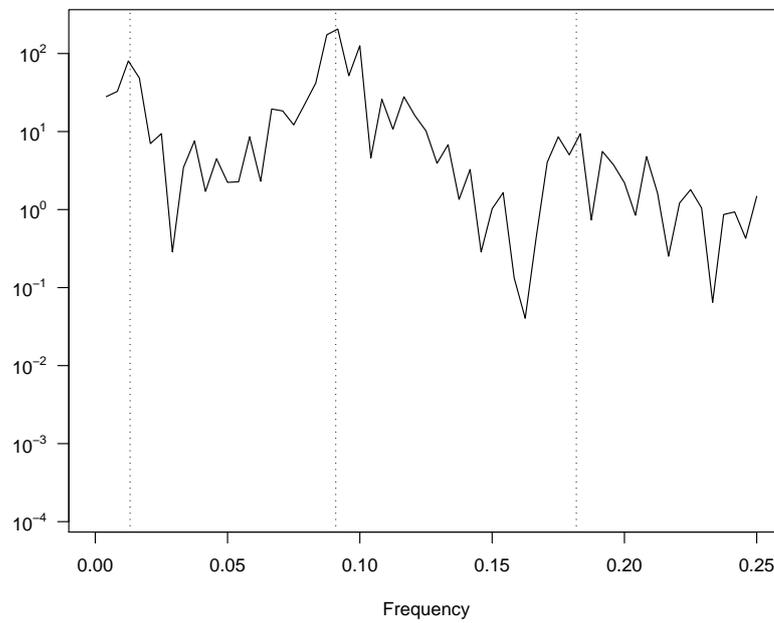


Figure 6.4: An expanded view of the the log periodogram computed for the square-root of the sunspot index series.

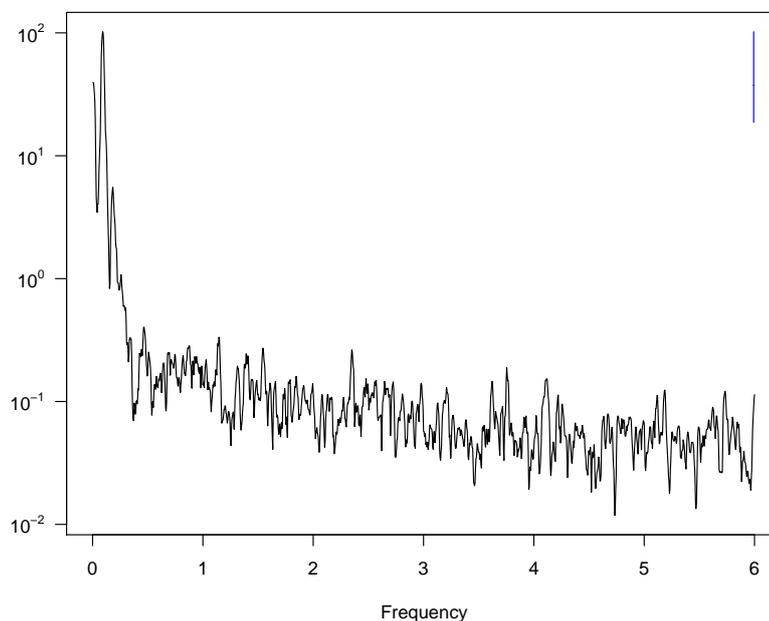


Figure 6.5: The log spectrum computed for the square-root of the sunspot index series.

is flat, it is appropriate to specify a large amount of smoothing. Where there are peaks or troughs a much smaller amount of smoothing should be used.

Figure 6.5 shows an estimate of the power spectrum for the sunspot series computed by averaging 7 periodogram values. A comparison with figure 6.3 shows that the variability of the estimate has been reduced a great deal. On the other hand, figure 6.6 shows that peaks visible in figure 6.4 have been smoothed out, making it harder to identify the corresponding frequencies. The location of the lowest peak is difficult to determine.

6.4.5 Tapering and Prewhitening

Power spectrum estimates obtained by averaging periodogram values can be improved in a number of ways. First, the process of sampling introduces discontinuities at the ends of the time series (a sudden drop to zero). The presence of these discontinuities introduces ripples into the periodogram and hence into power spectrum estimates. The effect of the discontinuity can be reduced by *tapering* the time series. This means at each end of the series the values are reduced toward zero by multiplying by a taper function.

One common choice for the taper function is the raised cosine. If the first K values are tapered, the effect is to multiply $X(k)$ by

$$\frac{1 - \cos(k\pi/(K + 1))}{2} \quad \text{for } k = 0, \dots, K - 1.$$

The effect of the tapering the last K values is defined similarly.

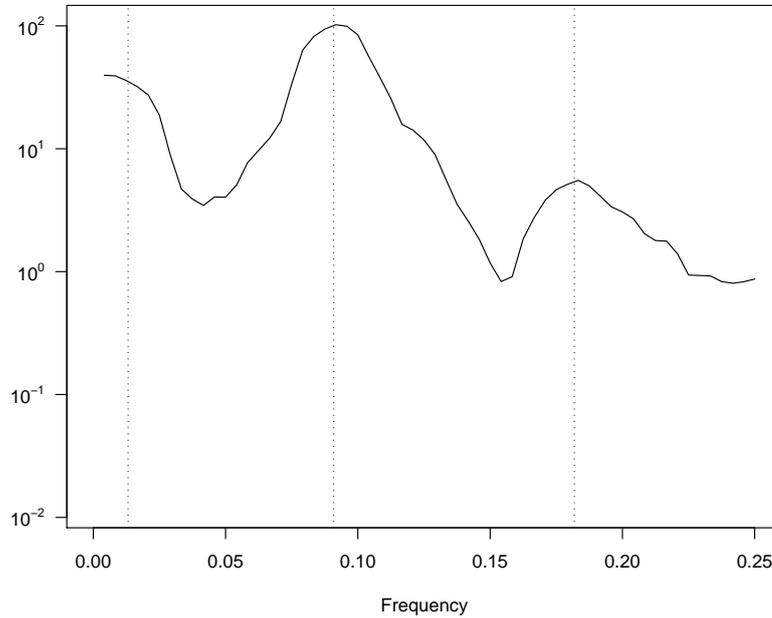


Figure 6.6: An expanded view of the the log spectrum computed for the square-root of the sunspot index series.

A second improvement which can be made in spectrum estimation is to use *pre-whitening*. This means removing obvious structure from the time series by filtering it before the spectrum is estimated. After the spectrum is estimated, it is *recoloured* using the transfer function of the filter, in a process known as *recolouring*. This process is useful when a very accurate estimate of the spectrum is required for some purpose.

6.4.6 Cross Spectral Analysis

Although it can be informative to look at the power spectrum for a single time series, it can be much more useful to use frequency domain methods to examine the relationship between series. The most important problem is that of making inferences about linear, time-invariant relationships of the form

$$Y(t) = \sum_{u=-\infty}^{\infty} a(u)X(t-u).$$

The information about the linear dependence between the time series $X(t)$ and $Y(t)$ is carried by the *cross-covariance function*

$$c_{YX}(u) = \text{cov}(Y(t+u), X(t)).$$

Provided that $X(t)$ and $Y(t)$ have summable autocovariance functions we can define the equivalent *cross-spectrum* in the frequency domain.

$$f_{YX}(\lambda) = \sum_{u=-\infty}^{\infty} c_{YX}(u)e^{-i\lambda u}$$

Estimation of the cross-spectrum is based on the *cross-periodogram*.

$$I_{YX}(\lambda)^T = \frac{1}{2\pi T} d_Y^T(\lambda) \overline{d_X^T(\lambda)}$$

In parallel with the result for power spectra, it is possible to show that

$$E\left(I_{YX}(\lambda)^T\right) \rightarrow f_{YX}(\lambda).$$

Again, in parallel with the power spectrum case, the cross-periodogram is an inconsistent estimator of the cross-spectrum, but consistent estimates can be obtained by smoothing. If B_T is chosen as in the power spectrum case then the estimator $f_{YX}^T(\lambda_0)$ obtained by averaging the cross-periodogram over an interval of width B_T centred on λ_0 provides a consistent estimator of $f_{YX}(\lambda_0)$.

$$f_{YX}^T(\lambda_0) \rightarrow f_{YX}(\lambda_0) \quad \text{as } T \rightarrow \infty.$$

Having defined the necessary parameters, we can now set down a simple *regression* model which relates two time series, namely

$$Y(t) = \sum_{u=-\infty}^{\infty} a(u)X(t-u) + \varepsilon(t). \quad (6.5)$$

Here $\varepsilon(t)$ is a stationary time series which is independent of $X(t)$ (note that we do not require that $\varepsilon(t)$ be white noise).

Now we note that

$$Y(t+u) = \sum_{v=-\infty}^{\infty} a(v)X(t+u-v) + \varepsilon(t+u)$$

and hence

$$Y(t+u)X(t) = \sum_{v=-\infty}^{\infty} a(v)X(t+u-v)X(t) + \varepsilon(t+u)X(t).$$

Taking expectations through this, we see that

$$c_{YX}(u) = \sum_{v=-\infty}^{\infty} a(v)c_{XX}(v-u).$$

Transforming into the frequency domain, we find that

$$\begin{aligned} f_{YX}(\lambda) &= \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} a(v)c_{XX}(v-u)e^{-i\lambda u} \\ &= \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} a(v)c_{XX}(v-u)e^{-i\lambda v} e^{-i\lambda(u-v)} \\ &= \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} c_{XX}(u)e^{-i\lambda u} \sum_{v=-\infty}^{\infty} a(v)e^{-i\lambda v} \\ &= f_{XX}(\lambda)A(\lambda) \end{aligned}$$

This suggests that we can estimate $A(\lambda)$ by

$$\hat{A}(\lambda) = \frac{f_{YX}^T(\lambda)}{f_{XX}^T(\lambda)}.$$

Taking the gain and phase of $\hat{A}(\lambda)$ will give us estimates of the gain and phase of $A(\lambda)$.

The independence of $X(t)$ and $\varepsilon(t)$ in equation 6.5 means that we have the following relationship between power spectra.

$$f_{YY}(\lambda) = |A(\lambda)|^2 f_{XX}(\lambda) + f_{\varepsilon\varepsilon}(\lambda)$$

where $f_{\varepsilon\varepsilon}(\lambda)$, the power spectrum of $\varepsilon(t)$, is known as the *residual spectrum*. The last equation can be written in the form

$$\begin{aligned} f_{YY}(\lambda) &= \frac{|f_{YX}(\lambda)|^2}{f_{XX}(\lambda)^2} f_{XX}(\lambda) + f_{\varepsilon\varepsilon}(\lambda) \\ &= \frac{|f_{YX}(\lambda)|^2}{f_{XX}(\lambda)} + f_{\varepsilon\varepsilon}(\lambda) \end{aligned}$$

which allows us to set down the following estimate for the residual spectrum.

$$f_{\varepsilon\varepsilon}^T(\lambda) = f_{YY}^T(\lambda) - \frac{|f_{YX}^T(\lambda)|^2}{f_{XX}^T(\lambda)}$$

To assess the quality of the fit of the model given by equation 6.5, we can compare the spectrum of the fitted series with that of the observed series.

$$\frac{|A(\lambda)|^2 f_{XX}(\lambda)}{f_{YY}(\lambda)} = \frac{|f_{YX}(\lambda)|^2}{f_{YY}(\lambda) f_{XX}(\lambda)}$$

The quantity

$$|R_{YX}(\lambda)|^2 = \frac{|f_{YX}(\lambda)|^2}{f_{YY}(\lambda) f_{XX}(\lambda)}$$

is called the *coherence* of the time series $Y(t)$ and $X(t)$ at frequency λ . $|R_{YX}(\lambda)|^2$ provides a measure of how related the frequency λ components of $X(t)$ and $Y(t)$ are. Its value lies between zero and one. An obvious estimate of $|R_{YX}(\lambda)|^2$ is obtained by substituting suitable power spectrum and cross-spectrum estimates into the last equation.

While the estimated gain and phase of the fitted filter give us complete information on the nature of the fitted relationship, it can also be useful to examine the estimated filter coefficients. The coefficients are collectively known as the *impulse response* of the filter because it is the result of applying a filter with coefficients $\{a(u)\}$ to the *impulse* series

$$X(t) = \begin{cases} 1 & t = 0, \\ 0 & \text{otherwise} \end{cases}$$

is to produce the output (or response) $a(t)$. The impulse response function can be estimated by direct inversion of the estimated transfer function.

Finally, it is possible (but not easy) to derive the asymptotic distributions of all the estimates presented in this section. The details of these derivations and the generalisation to multivariate time series can be found in David Brillinger's book *Time Series: Data Analysis and Theory*.

6.5 Computation

6.5.1 A Simple Spectral Analysis Package for R

There is a simple spectral analysis package I put together because the current facilities available in R and S are not as good as they could be. You can pick up the R code from

```
http://www.stat.auckland.ac.nz/~ihaka/726/spectrum.R
```

Use the `source` command to read this into R.

The code is not particularly sophisticated, but should give you some idea of what is possible. You should also be able to see that it corresponds directly to formulae you've seen in class. (I am planning to implement a real spectral analysis library in the future).

Note that all frequencies (including the smoothing bandwidths) are specified in "cycles per unit time". E.g. for monthly data, a frequency of 1/12 corresponds to a yearly cycle.

6.5.2 Power Spectrum Estimation

These calls are appropriate for estimating the power spectrum of a single series. More sophisticated techniques are possible, but this will get you most of the information present in a series.

1. Compute and plot the periodogram.

```
z = power.spectrum(x)
plot(fxx(z))
```

2. Compute and plot a smoothed periodogram estimate with a given smoothing bandwidth.

```
z = power.spectrum(x, bw = .01)
plot(fxx(z))
```

3. Compute and plot a smoothed periodogram estimate with a given smoothing span. The span is an odd, positive integer and the estimate will be obtained by averaging `span` periodogram ordinates.

```
z = power.spectrum(x, span = 11)
plot(fxx(z))
```

4. There are a number of options to the plotting command. Zoom in on a given frequency range.

```
plot(fxx(z), xlim = c(0, .2))
```

Plot the confidence limits around the estimate.

```
plot(fxx(z), ci = TRUE)
```

6.5.3 Cross-Spectral Analysis

The model to be fitted is

$$Y(t) = \sum_{u=-\infty}^{\infty} a(u)X(t-u) + \varepsilon(t)$$

The key assumption is that $\varepsilon(t)$ be a stationary, mixing series. $X(t)$ and $Y(t)$ are not required to be stationary, although this is case we have theory for.

1. Estimate the Spectra and Cross Spectra (no plotting).

```
z = spectrum(x, y, bw = .01)
```

2. Display the spectra for the X and Y series. The `ci` and `xlim` options can be specified.

```
plot(fxx(z))
plot(fyy(z))
```

3. Display the coherence (the 95% null point is shown).

```
plot(coherence(z))
```

4. Display the spectrum of the residual series (`ci` and `xlim` also apply in this case).

```
plot(residual.spectrum(z))
```

5. Display the gain and phase of the fitted filter (`ci` and `xlim` also apply in this case).

```
plot(gain(z))
plot(phase(z))
```

6. Display the filter coefficients (the impulse response).

```
plot(impulse.response(z))
```

6.6 Examples

The series `berlin` and `vienna` (available from the class web site) contain the monthly average temperatures (in °C) observed in Berlin and Vienna over the years 1775–1950. We will examine these series using the cross-spectral techniques developed in section 6.4.6. We will take the Vienna series to be $X(t)$ and the Berlin series to be $Y(t)$ and we will fit a model of the form

$$Y(t) = \sum_{u=-\infty}^{\infty} a(u)X(t-u) + \varepsilon(t).$$

(Note that we are doing this as a pure exploration, not because we think that there is any kind of causal influence of Vienna temperatures on Berlin ones.)

We will begin by trying to obtain a reasonable estimate of the spectra of the x and y series. It is usual to have to experiment with various amounts of

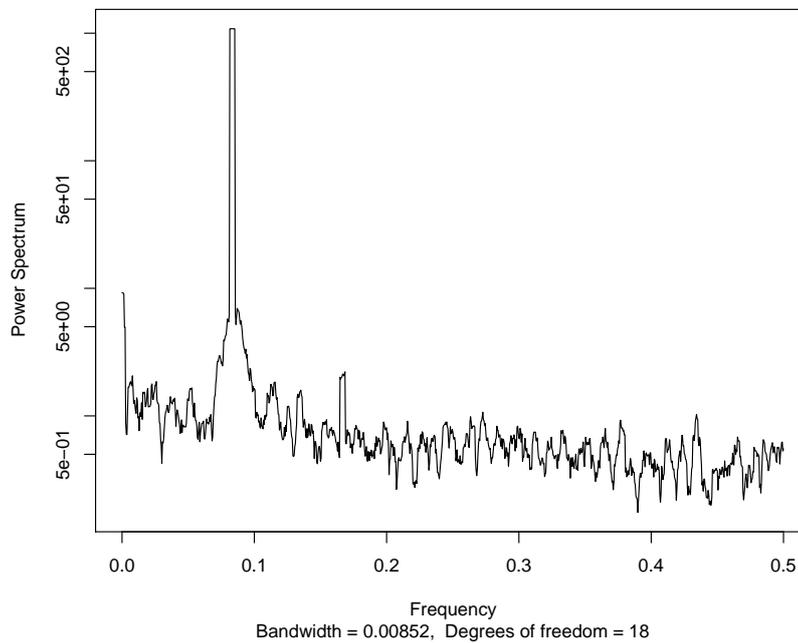


Figure 6.7: The spectrum of the Vienna temperature series.

smoothing when obtaining spectrum estimates. We'll begin with just a small amount of smoothing. This will enable us to look precisely at any periodic components of the series. The first estimate smooths 9 periodogram estimates.

```
> z = spectrum(vienna, berlin, span=9)
> plot(fxx(z))
```

Figure 6.7 shows the resulting power spectrum. The highest peak is at frequency $1/12$ and corresponds to the yearly seasonal cycle. The peak to the right of this is at frequency $2/12$ and so is a harmonic which affects the shape of seasonal waveform. The peak very close to frequency zero indicates that there is some type of long period variation or trend in the series.

Expanding the low frequency region of the spectrum enables us to get a better look at the most interesting part of the spectrum. This can be done as follows:

```
> plot(fxx(z), xlim = c(0, 0.2))
```

The result is shown in figure 6.8. No additional information is revealed.

We can also extract the power spectrum for the Berlin series and examine it in the same way using the command

```
> plot(fyy(z), xlim = c(0, 0.2))
```

The result is shown in figure 6.9. The same major seasonal peak appears in the Berlin spectrum, but the harmonic at frequency $2/12$ is diminished. This suggests that the temperature variation in Berlin is closer to being a pure sine

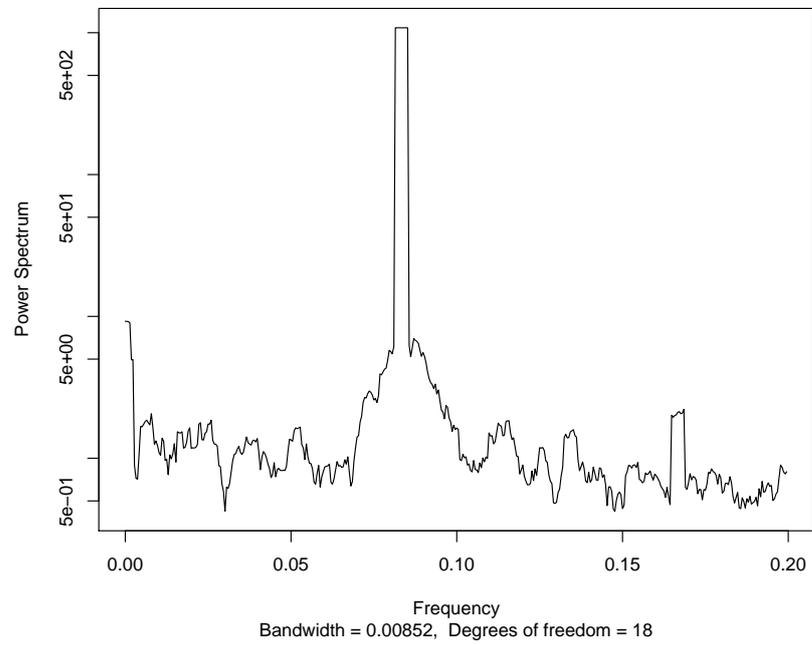


Figure 6.8: The spectrum of the Vienna temperature series.

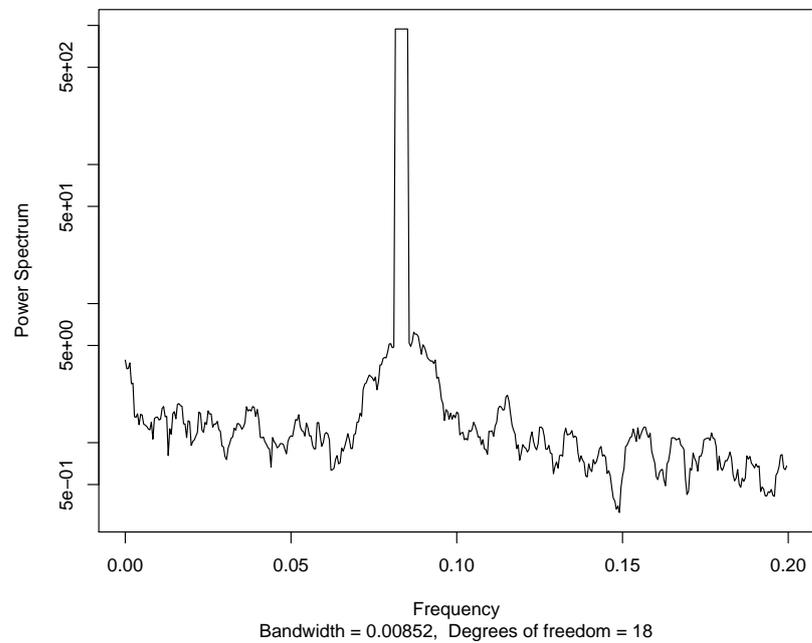


Figure 6.9: The spectrum of the Berlin temperature series.

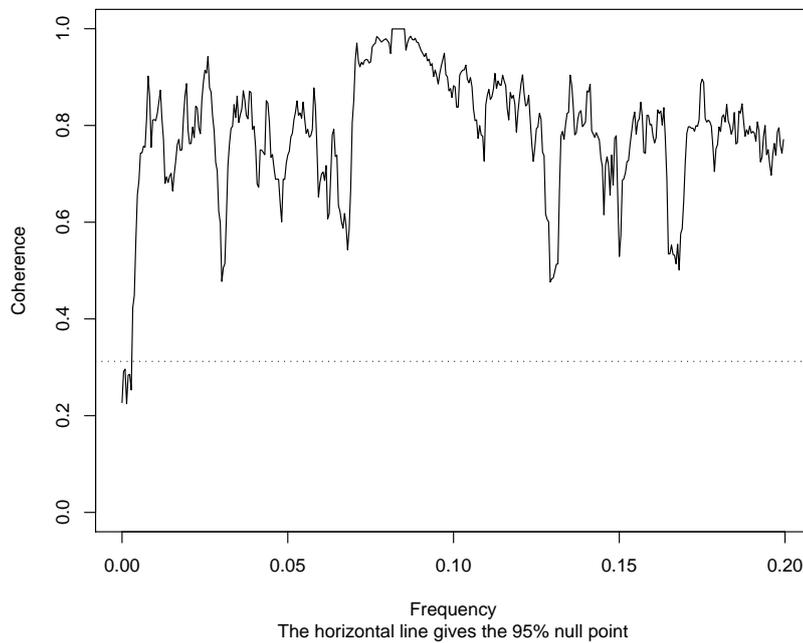


Figure 6.10: The coherence of the Vienna and Berlin temperature series.

wave than it is in Vienna. In addition, the low frequency portion of the Berlin spectrum is lower than that of the Vienna one.

We begin an examination of the relationship between the two series by looking at the coherence between them. This can be done with the command

```
> plot(coherence(z), xlim = c(0, 0.2))
```

The coherence is high right across the spectrum although it is clearly strongest in the region of the spectral peak at frequency $1/12$. There is also a region of low coherence in the very lowest frequency region.

The message that this plot conveys is that the two temperature series are strongly related. We can examine the best fitting filter in either the frequency domain or the time domain. We'll begin in the frequency domain. The gain of the best fitting filter can be plotted with the command

```
> plot(gain(z))
```

The result is shown in figure 6.11. The figure shows that that the estimated gain varies about a horizontal line which is slightly less than 1 in magnitude. This means that the temperatures in Berlin show a similar pattern to those in Vienna, but are slightly colder. The exception to this general rule is that the long-term slow variations appear to be unrelated. This is in agreement with what the coherence shows.

The alternative way of looking at the filter is to examine the filter coefficients or impulse response. A plot of the impulse response can be produced with the command

```
> plot(impulse.response(z))
```

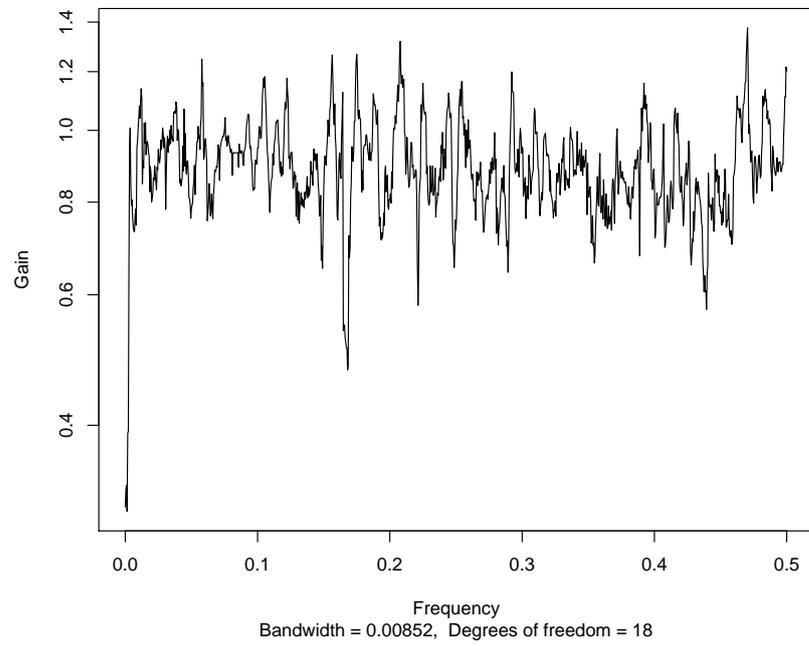


Figure 6.11: The gain of the filter which approximates the Berlin series using the Vienna one.

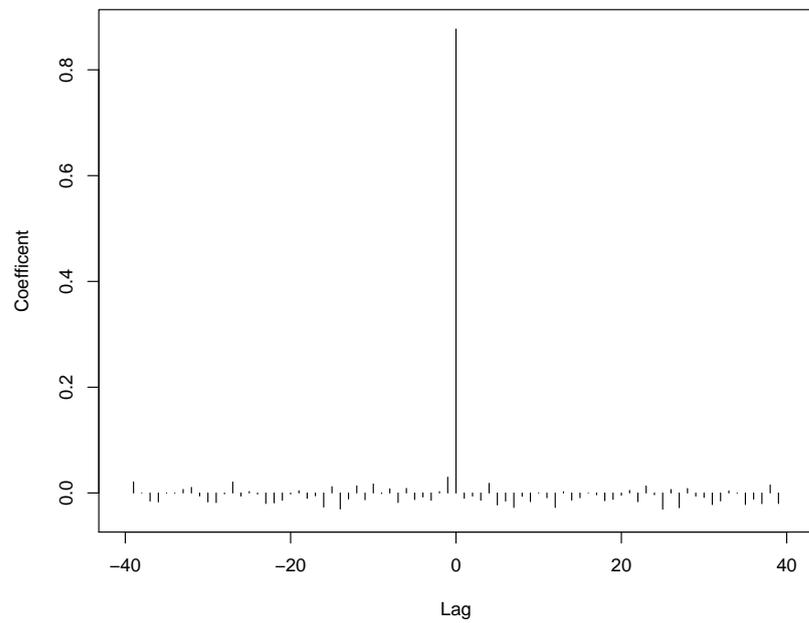


Figure 6.12: The coefficients of the filter which best approximates the Berlin series using the Vienna one.

This produces the result shown in figure 6.12. The plot shows that the average monthly temperatures in Berlin and Vienna move in step, with the temperature in Berlin being between 0.8 and 0.9 times that in Vienna.

The cities of Berlin and Vienna are geographically close and weather systems pass between them in a matter of days. Because of this it should be no surprise that the best fitting filter has the form that it does. What is surprising is that there is such a lack of similarity at the very lowest frequencies. This suggests that long term temperature variations are a local effect, possibly the result of human activity.