Serie documentos de trabajo

APPLIED TIME SERIES ANALYSIS

Moguens Bladt

DOCUMENTO DE TRABAJO

Núm. VI - 1995

Contents

Preface

The material of these lecture notes was originally presented in a graduate econometrics course at El Colegio de Mexico, Mexico City. Emphasis has been put on applications, and on theory that is important in practice. This means that proofs have in many places been replaced by discussions of the main ideas and their relation to pratrical applications.

It is the intention of the notes that they will enable students to perform a statistical analysis of times series data using advanced modern techniques.

A variety of statistical packages are available today, and most of them provide time series analysis at some level. In these notes we use the package ITSM (Interactive Time Series Modelling; Brockwell and Davis, 1991, Springer Verlag) as a reference.

The notes are organized as follows. In chapter 1 we give a description of stationarity, transformations, trends removals and differencing techniques. Chapter 2 accounts for a detailed treatment of ARMA models, and generalized linear processes. In chapter 3 we describe various estimation procedures, both parametric and non–parametric, order selection techniques and goodness of fit procedures. Multiplicative models and (S)ARIMA models are dealt with in chapter 4, and in chapter 5 we introduce the important frequency analysis. Multivariate time series and transfer function models, together with multivariate spectral analysis, are presented in chapter 6.

The reader may wonder why prediction theory has not been presented. The reason is that prediction of ARMA models, and related models, using best linear predictors, is a trivial matter once the analysis has been performed. Prediction theory is, however, far from trivial, and the classical time series method is only one way of approaching the problem. A more profound introduction to forecasting would involved a whole course.

The exposition has been kept reference free and self–contained. At the end the reader will find a brief bibliographic guide for further reading.

Mexico City
June, 1995

Mogens Bladt

Chapter 1

Stationarity and Transformations

1.1 Introduction

The assumption of stationarity is essential in most applications of time series analysis. The reason for this is that unlike classical statistical problems we have usually only one single oberservation available, which is the realization of the whole series itself.

Statistical methods are based on replications of experiments, so that we are provided with information about an experiment using various measurements. In the standard theory these measurements are assumed to be independent identically distributed (i.i.d), and the interpretation is simply that the experiment is performed a number of times independently of each other. These measurements constitutes the basis of the statistical analysis.

In time series analysis we observe observations through time, and usually we are only provided with one observations per time unit. This could for example be the Dow Jones index every day, the volume traded at Wall Street every week or an exchange rate at a certain time of the day. In all these examples replication is not possible because of the nature of the data. Even more, the quantities in these examples are purely deterministic, and can hence be recorded without error.

The goal of time series analysis is to describe data in such a way that it may provide an understanding of the phenomena we are considering, or to extract some basic features from the data which enable us to make predictions. To this end we would like to extract statistical measures from the data, such as the correlation between variables. The basic assumption of stationarity means that the correlation from one day to the other, from one week to the next etc., does not depend on which days or weeks we consider: if we for example consider the a measurement of the volume traded per week at Wall Street through one year, the correlation between traded volumes at week 1 and 2 is the same as between week 34 and 35.

Notation 1.1.1 *Throughout the notes we will use capital letters for random variables, such as X, Y, X_5 etc., and small letters for the corresponding data/observations, such as x, y, x_5.*

Let $..., X_{-n}, ..., X_{-1}, X_0, X_1, ..., X_n, ...$ be a time series. We shall use the notation $X = \{X_t\}_{t\in\mathbb{Z}}$. Then we have

Definition 1.1.1 *X is (second order) stationary if $\mathbb{E}(X_t)$ does not depend on t, $\mathbb{E}(X_t^2) < \infty$ and*

$$\mathbf{Cov}(X_t, X_{t+h}) = \gamma(h)$$

only depends on h, for all $t, h \in \mathbb{Z}$. In that case, $\gamma(h)$ is called the auto-covariance function of $\{X_t\}$. Introducing the variance of series

$$\sigma_X^2 = \mathbf{var}(X_t) = \gamma(0),$$

we define the auto-correlation function $\rho(h)$ by

$$\rho(h) = \frac{\gamma(h)}{\sigma_X^2} = \frac{\gamma(h)}{\gamma(0)}.$$

Note: The second assumption $\mathbb{E}(X_t^2) < \infty$ is only technical, and needed for the variance to exists, so that the covariance exists as well.

In a time series the most important feature to exploit is the correlation between the measurements at different times. It is the correlation that tells us to which degree the variables are related to each other, and hence gives important information to be used when predicting future values.

For an observed time series it is therefore desirable to estimate the correlation or covariance between the variables. If we would not have assumed stationarity, but simply allowed the correlation between data to vary arbitrarily with time, we would not be able to estimate the correlation between the data values. To illustrate this point consider a time series with observations $x_0, x_1, ..., x_N$. Assume that series has zero mean. We want to estimate the correlation between any two consequtive point, X_0 and X_1, X_1 and X_2 etc. If we do not assume stationarity we simply have that

$$\begin{aligned} \hat{\mathbf{Cov}}(X_0, X_1) &= x_0 x_1 \\ \hat{\mathbf{Cov}}(X_1, X_2) &= x_1 x_2 \\ \cdots &\quad \cdots \end{aligned}$$

which are useless estimates since they are only based on one observation, namely the empirical covariance between x_0 and x_1. If we in turn assume stationarity we have that

$$\mathbf{Cov}(X_t, X_{t+1})$$

is the same for all t, and hence we can estimate the covariance as the average

$$\hat{\mathbf{Cov}}(X_t, X_{t+1}) = \frac{1}{N+1}\sum_{i=0}^{N} x_i x_{i+1}.$$

This estimate is of course much better, and is in fact useful in an statistical analysis. Therefore stationarity is essential and important.

Many time series which are obviously not stationary are hence transformed into stationary time series, and statistical analysis are performed to the stationary series. When conclusions are drawn or prediction performed, the stationary series is then transformed back to its original.

1.2 Formal framework

In this section we introduce carefully terms and definitions, some of which have already been introduced vaguely in the introduction and which are more or less assumed to be well known from basic statistics.

For a random variable X we denote its mean or expected value by $\mathbb{E}X$ or $\mathbb{E}(X)$. The covariance between two random variables X and Y $\mathbf{Cov}(X, Y)$ is given by

$$\mathbf{Cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\right).$$

In particular, the variance of X is

$$\mathbf{Var}(X) = \mathbf{Cov}(X, X) = \mathbb{E}\left((X - \mathbb{E}X)^2\right).$$

A time series is a collection of random variables $\{X_t\}_{t\in\mathbb{Z}}$ indexed by the integer numbers $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$. That we index by both negative and positive numbers has no practical implications, since any sample we may consider is finite and usually indexed only by positive numbers. The reason, however, is merely mathematical convenience.

There exists actually various kinds of stationarity for time series. One has been imported from probability theory is called strict stationarity, and means that distributions of a series $X = \{X_t\}_{t\in\mathbb{Z}}$ and its translated series $X_h = \{X_{t+h}\}_{t\in\mathbb{Z}}$ are the same for all $h \in \mathbb{Z}$.

Another stationarity criterion, which we shall use, is called second order stationarity, and has been introduced in the introduction. The reason for the term 'second order' is that the criterion is based on only the first two moments of the distribution, namely the expectation (mean) and the variance/covariance.

Second order stationarity and strict stationarity are equivalent whenever the time series under consideration are Guassian, that is any sub–vector $(X_{t_1}, X_{t_2}, X_{t_3}, ..., X_{t_m})$ of time series values has a multivariate normal distribution. This implies for example that all elements in the time series X_t are normally distributed, and any linear combination of terms from the series as well.

1.3 Stationary Time Series in practice: some data sets

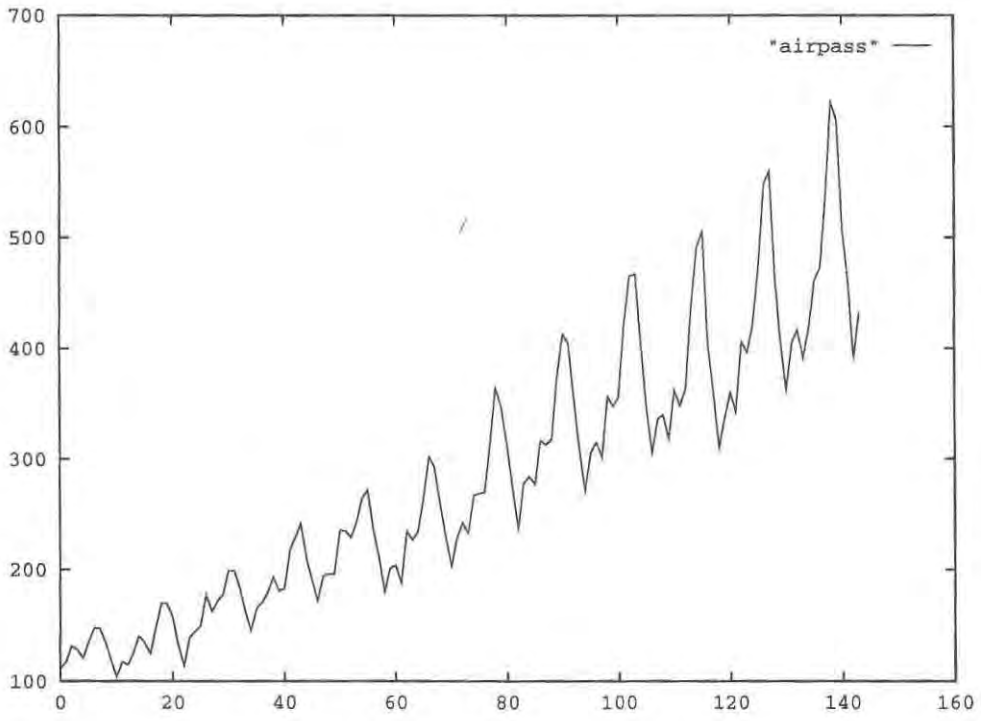

Figure 1: Airpass data

Since stationarity is a condition under which we can perform a statistical analysis it is crucial that we learn how to distinguish stationarity from non-stationarity. There exists in the literature no test for stationarity, that can compete with such basic tools as visual inspection of the time series and its auto–correlation function.

There exists mainly two kinds of viaolation to stationarity. The first is the presence of a trend, while the other is the presence of a certain kind of periodicities, often called

seasons or seasonality.

We will consider some examples of times series that have occurred in practice.

The Airpass data set (see Figure 1) shows international airline passenger totals in thousends from January 1949 to December 1960. The data set shows various important features. First of all it cannot be stationary since there is an increasing trend. Furthermore, cycles or periods of length 12 is present, and finally the sizes/hights of the cycles increases with increasing time.

All the three features makes it non–stationary. The periods means that e.g. january and february data are not correlated in the same way as november and december (in the first case there is an increase and in the latter a decrease). However, if we consider only janauries to februaries through the whole period we may find the correlations of quite similar nature.

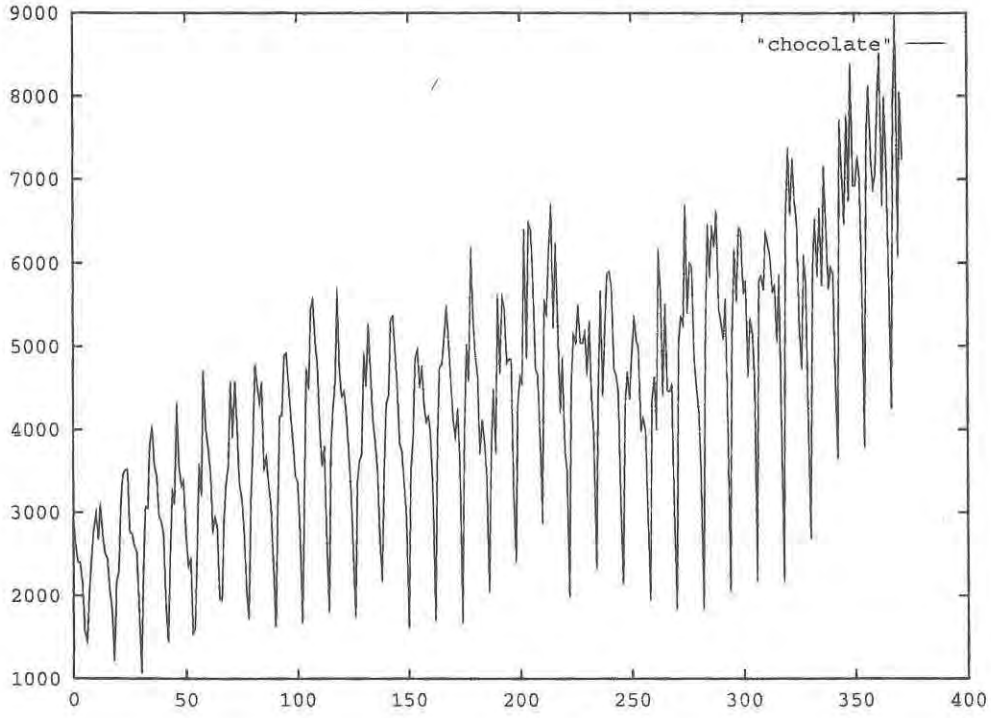

Figure 2: Chocolate data

The increase in the hights of the periods makes the time series behave wilder and wilder as time goes, and this is clearly a viaolation to the stationarity idea that variations do not depend on time. Therefore it is desireable to transform the data in such a way that we equalize the hights of the periods.

Thirdly, the increasing trend of the data makes them clearly non–stationary, since this implies that the means of the random variables through the years are increasing as well.

The chocolate data (Figure 2) shows a clearly non–stationary behavior (why?), and it will be the first task of an analysis to remove trends, cyclic behavior and equalize the size of the periodicities. The periodicities do not show too clearly an increasing or decreasing trend, so in this case it will be a matter of trial and error with Box-Cox transformations (see section 1.4) to stabilize the sizes of these cycles.

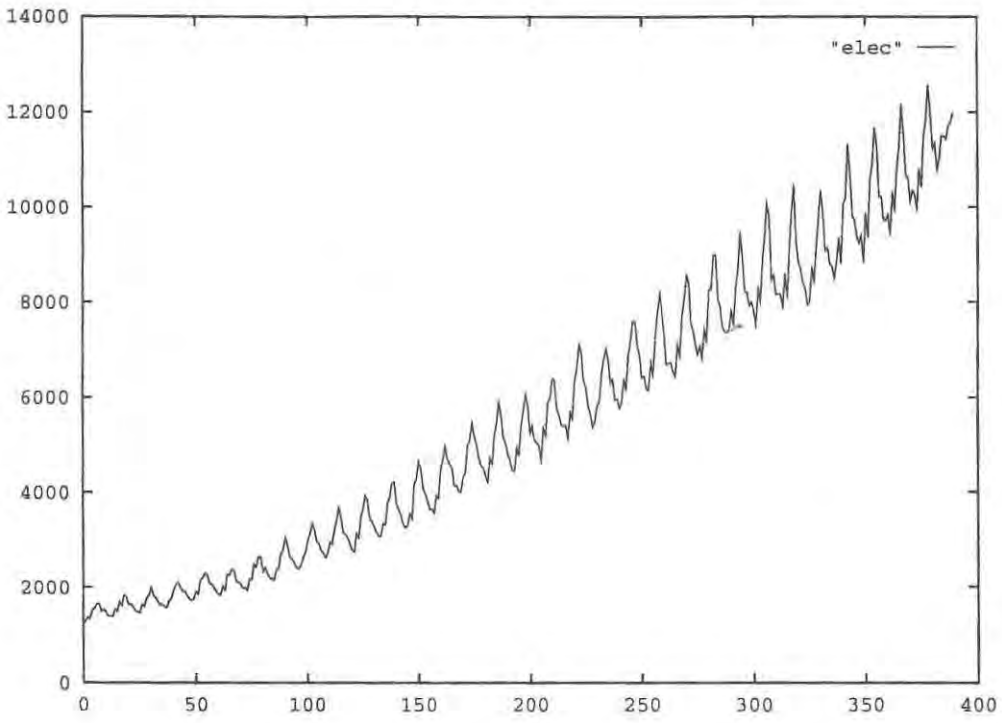

Figure 3: Electricity data

Also the overall increasing trend is not too clear, but a linear trend does not show succesful one might try a polynomial of order 3,4 or 5 to get the trend under control.

The cycles, however, seem to be well behaved and should not cause any trouble in removing.

The electricity data are in some senses similar to the airpass data, only that the overall increase looks quadratic rather than linear, at least at the beginning of the data set.

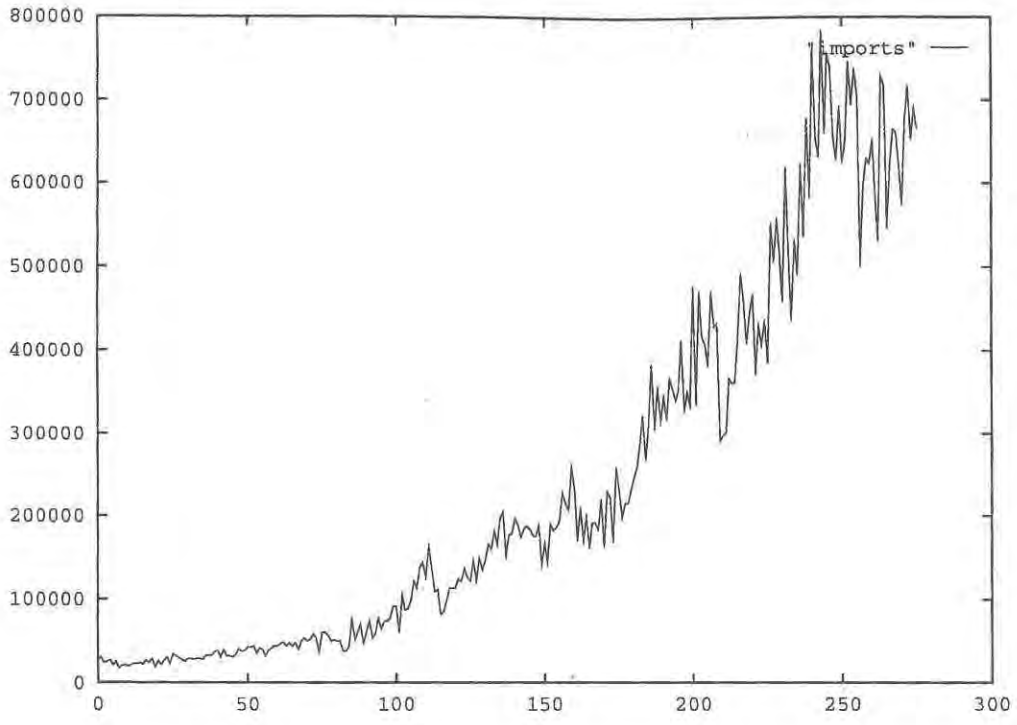

Figure 4: Imports data

This rather erratic time series has an overall increasing trend, but it is yet not clear how transform this series into an stationary one, if at all possible. A solution may be to treat is as a non–stationary ARIMA model, which we shall return to later in these notes.

1.4 Box–Cox transformations

The Box–Cox transformation can be characterized as a preliminary transformation. That is, the Box–Cox transformation should be applied before any serious transformations/analysis starts, and is usually the first thing to do if necessary.

The Box–Cox transformation is as follows. If $x_1, ..., x_N$ is our time series data, then the Box–Cox transformed datas are given by

$$y_i = f_\lambda(x_i) = \begin{cases} \frac{x_i^\lambda - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log(x_i), & if\ \lambda = 0. \end{cases}$$

This transformation is useful when the data shows an increasing or decreasing trend of variability. For example, if the standard deviation of the X_i's increases linearly for increasing i's, then the logarithmic transformation will make the variability constant.

It is really a matter of trial and error to finding the right Box–Cox transformation in special cases of application, but the methods applies whenever the data presents a trend, that is an increase or decrease, in the variability over time.

The Box-Cox transformation is only a special case of a more general theory of variance stabilizing methods. In general, if the variance of X_t is $\sigma^2(t)$, then one should apply the function to the original data given by

$$f(y) = \int_0^y \frac{c}{\sqrt{\sigma^2(t)}} dt,$$

where c is an arbitrary constant. Let us see what this means in the Box-Cox case. If $f(y) = \log(y)$ then we obviously have that

$$\log(y) = \int_0^y \frac{c}{\sqrt{\sigma^2(t)}} dt,$$

which then gives (differentiate both sides with respect to y) that

$$\sigma(t) = ct.$$

In case of $f(y) = \frac{y^\lambda - 1}{\lambda}$, $\lambda \neq 0$, we have the equation

$$\frac{y^\lambda - 1}{\lambda} = \int_0^y \frac{c}{\sqrt{\sigma^2(t)}} dt,$$

which in a similar way gives that

$$\sigma(t) = ct^{1-\lambda}.$$

Thus we note the following: $\lambda = 0$ corresponds to a linear increase in the standard deviation, $\lambda \in (0, 1]$ to a concave increase, $\lambda \in (-\infty, 0)$ to a convex increase, $\lambda \in (1, 2]$ to a slow decrease and $\lambda \in (2, \infty)$ to a fast ploynomial decrease.

Obviously we can come up with much more sophisticated choises of f, but the Box-Cox has become standard in time series literature, and has proved a good flexibility compared to its simplicity.

1.5 Classical trend removal techniques

The general model for a time series that has a certain trend m_t, say, and a seasonal component s_t is given by

$$X_t = Y_t + m_t + s_t,$$

where Y_t is a stationary time series. The point is then to remove the trend m_t and the seasonality s_t from our originally data X_t such that we are left with a stationary series Y_t.

If there were no seasonality present we have the model:

$$X_t = Y_t + m_t.$$

Assume without loss of generality that Y_t has zero mean. The mean can always be put into the trend term if necessary. Attempts has been made in the literature to consider m_t as a member of a parametric family of functions, e.g.

$$m_t = a_0 + a_1\ t + a_2\ t^2 + ... + a_n\ t^n$$

for some fixed n. For $n = 1$ we have a linear trend. A method to estimate the coefficients $a_0, a_1, ..., a_n$ would be to minimize

$$\sum_i (x_i - m_i)^2$$

with respect to $a_0, a_1, ..., a_n$ using a least squares method. Then subtracting the estimated

$$\hat{m}_t = \hat{a}_0 + \hat{a}_1\ t + ... + \hat{a}_n\ t^n$$

from X_t should yield a stationary sequence $\{Y_t\}$.

The method above can be considered as a parametric estimation of m_t. We could also choose to use a non–parametric method, using only the data themselves to estimate m_t. For example one could choose to smooth the data using a moving average procedure like

$$\hat{m}_t = \frac{1}{2q+1} \sum_{i=-q}^{q} x_{t+i},$$

for some choice of q. When is this estimate good? Our data are given by the model

$$X_t = Y_t + m_t$$

so

$$\frac{1}{2q+1} \sum_{i=-q}^{q} x_{t+i} = \frac{1}{2q+1} \sum_{i=-q}^{q} y_{t+i} + \frac{1}{2q+1} \sum_{i=-q}^{q} m_{t+i}.$$

Thus if e.g. m_t is assumed to be constant, linear or symmetrical over $[t-q, t+q]$ then the first sum equals m_t approximately, since the average of the 'noise' terms can be assumed to be approximately zero due to the law of large numbers.

Main advantage of the parametric method is that we are provided with a function that is known beyond the observed time scale, so that prediccion does not become a problem. The main disadvantage is that it only provide us with little flexibility in the number of possible trends, since we have to keep the number of parameters fairly low for ordinary applications (less than 3 e.g.).

The main advantage of the non–parametric method is that it will almost surely remove the trend within the oberserved time span, but it does not provide us with any estimate or information obout the possible trend in the future. Thus the method is more descriptive than of possible use for the sake of predictions.

The moving average smoothing is by the way a special case of filtering a time series: we say that $\hat{m}_t$ has been obtained using a linear filter to x_t. We shall return to this subject in more detail later.

For the general case where both seasonal components and trend is present we can do a similar thing. If the period is d, say, then we can smooth our data by applying a linear filter as above. If the d is an unequal number, $d = 2q + 1$, then we estimate the trend as

$$\hat{m}_t = \frac{1}{d} \sum_{i=-q}^{q} x_{t+i}.$$

If $d = 2q$ is equal, then we have a slight problem, because in the sum $\sum_{i=-q}^{q} x_{t+i}$ there always appear an unequal number of terms, and we only want to smooth over the period d. This problem is overcomed by only adding half weight to the end terms in the sum, i.e.

$$\hat{m}_t = \frac{1}{d}\left(\frac{1}{2}x_{t-q} + x_{t-q+1} + ... + x_{t+q-1} + \frac{1}{2}x_{t+q}\right).$$

Once the trend has been estimated, we estimate the seasonal effect by removing the trend from the original series. The seasonality effect is of course the same for all cycles, by definition, and hence we can estimate it by an average

$$w_k = \frac{1}{[N/d]} \sum_j \left(x_{k+jd} - \hat{m}_{k+jd}\right), k = 1, 2, ..., d$$

where j is such that $q < k + jd \leq n - q$, and $[a]$ denotes the integer part of a. Now this average does not necessarily sum to zero (a desirable property), so we will use the estimate

$$\hat{s}_k = w_k - \frac{1}{d} \sum_{i=1}^{d} w_k.$$

Extend $\hat{s}_k$ to all k by $\hat{s}_k = \hat{s}_{k-d}$ for $k > d$. Now remove the seasons from the data, to obtain deseasonlized data

$$d_t = x_t - \hat{s}_t$$

and reestimate the trend of the data, either by smoothing/moving average or by a polynomial fit. Finally remove this trend from the data to obtain the stationary series.

1.6 Differencing

The classical approach to trend removal and deseasonalization may seem reasonable in practice, though it may look rather ad hoc orientated from a scientific point of view. Introducing differencing as an alternative tool we will be able to remove both seasonality and any polynomial trend at any degree by one single method. Define the backward shift operator B by

$$BX_t = X_{t-1}.$$

We define powers of B in the usual way

$$B^j = B\left(B^{j-1}X_t\right) = ... = X_{t-j}.$$

Define the difference operator by

$$\nabla = 1 - B.$$

Then $\nabla X_t = X_t - X_{t-1}$. Powers are defined as above $\nabla^j = (1-B)^j$ (not to confuse with $1 - B^j$).

With this difference operator at hand we will be able to remove seasonality as well as polynomial trends. Indeed consider the general model

$$X_t = Y_t + m_t + s_t,$$

where s_t is a seasonality component with period d, say. Let us start with considering a differencing of lag d. Due to seasonality we obviously have that $\nabla_d s_t = 0$, where ∇_d is the lag–d difference operator, $\nabla_d = 1 - B^d$, $\nabla_d X_t = X_t - X_{t-d}$. Thus we have

$$\nabla_d X_t = m_t - m_{t-d} + Y_t - Y_{t-d}.$$

In this differenced model the trend is now $m_t - m_{t-d}$ and the (stationary!) noise term is $Y_t - Y_{t-d}$. Moreover the noise term has mean zero. Thus the situation now is equivalent with a model of the type

$$X_t = Y_t + m_t$$

where Y_t is stationary with mean zero and m_t a trend. Assume that this trend is polynomial (if the original trend above is polynomial, then so is its differenced (deseasonlized) trend), so we may write

$$m_t = a_0 + a_1 t + a_2 t^2 + ... + a_k t^k.$$

Then apply ∇^k to this polynomial to get

$$\nabla^k m_t = k! a_k.$$

Thus ∇^k fully removes the polynomial trend of order k, and leaves back a constant term, which only adds to the mean. Then we have that

$$\nabla^k X_t = k! a_k + \nabla^k Y_t$$

where $\nabla^k Y_t$ is stationary. Thus X_t is stationary.

In particular, ∇ removes linear trends, ∇^2 quadratic terms and ∇^3 cubic trends. For more sophisticated trend removals, like logarithmic or exponential trends, we would have expand these functions into Taylor series and apply an approximating polynomial of not too high degree as a trend. If this does not seem feasible we must recommend the calssical approach for trend removal, since it is not desireable to use an exessive number differencing steps, because each step decreases the number of data points availabe for our analysis.

Chapter 2

ARMA processes

ARMA processes is the most famous class of time series in the literature, and that is mainly due to two reasons. Firstly they are mathematically relatively easy to deal with, and secondly it can be shown that any stationary time series we may be confronted with in practice can be approximated arbitrarily close by an ARMA model.

A generalization of ARMA processes, the so called generalized linear processes, will also be presented. The general linear processes will show that any ARMA process, under mild regularity conditions, can be represented as an infinite MA process. This fact will be used in model selection of ARMA processes, and is therefore as such an important statement to notice.

2.1 The simplest model: i.i.d.

The notation i.i.d. means independent, identically distributed. This is a standard assumption in other parts of statistical analysis, which does not appeal too much to time series analysis. Everything independent would simply imply that any future predictor would have to be generated as a random number independently of the past. This means that our observations only serve to settle the distributional properties, but any forecast would be independent as such of our observations.

In spite of this, the i.i.d. assumption is still an important one in time series analysis. In moving average models we are dealing with i.i.d.'s and in error terms as well.

Let $X_t, t \in \mathbb{Z}$ be i.i.d. with variance σ^2. Then the covariance function $\gamma(h)$ is given by

$$\gamma(h) = \mathbf{Cov}(X_t\ X_{t+h}) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

The terminology in time series states that any time series having auto–covariance function (2.1) is called a white noise. The reason for this apparently strange name will be

clear later from spectral analysis where we see that the socalled spectrum of a white noise process is constant. This means that all frequencies are present with the same intensity, and from physics it is known that this would imply white light.

2.2 AR(1) processes

Auto–regressive processes of first order, AR(1), are time series $\{X_t\}$ that have the property

$$X_t = aX_{t-1} + \epsilon_t,$$

where $\{\epsilon_t\}$ is a white noise process. In this model the value of the process at time t, X_t, is obtained through a simple linear regression on its own past values X_{t-1}. Thus a plays the role of the slope and the white noise plays the role of the error term. This is the reason for the name 'auto–regressive'.

The auto–regressive property implies that we can write

$$\begin{aligned} X_t &= aX_{t-1} + \epsilon_t \\ &= \epsilon_t + a\left(aX_{t-2} + \epsilon_{t-1}\right) \\ &= \epsilon_t + a\epsilon_{t-1} + a^2\left(aX_{t-2} + \epsilon_{t-2}\right) \\ \ldots &= \ldots \\ &= \epsilon_t + a\epsilon_{t-1} + a^2\epsilon_{t-2} + \ldots + a^{t-1}\epsilon_1 + a^t X_0. \end{aligned}$$

If we make the assumption that $X_0 = 0$, and that the white noise process has mean μ_ϵ and variance σ_ϵ^2, then

$$\mathbb{E}\left(X_t\right) = \mu_\epsilon\left(1 + a + a^2 + \ldots + a^{t-1}\right) = \begin{cases} \mu_\epsilon \frac{1-a^t}{1-a} & \text{if } a \neq 1 \\ t\mu_\epsilon & \text{if } a = 1 \end{cases}.$$

In this case we can see that X_t is not stationary, since its mean depends on t. This means that we cannot have a stationary process where we impose the condition $X_0 = 0$, or any other fixed value for that sake.

The reason for we did not obtain a stationary time series in this case is not only because we restricted X_0 to 0, but merely because we constructed the whole process as depedent on this particular value. The key point is that stationarity is a feature of processes that have been running for a very long time. That is why we have defined X_t for all $t \in \mathbb{Z}$, and we will be thinking of the process as comming from infinitely far, namely from $t = -\infty$. When the process then hits our usual region for observation $t = 0, 1, \ldots, N$, then the process has developed into stationary mode.

In the case where $X_0 = 0$ we can compensate for this assumption by letting $t \to \infty$. We see that if $|a| < 1$ then $\mathbb{E}X_t$ has a limit as $t \to \infty$, namely $\mu_\epsilon/(1-a)$.

For $h \geq 0$ we have that the covariance function is given by

$$\begin{aligned}
\mathbf{Cov}(X_t, X_{t+h}) &= \mathbf{Cov}\left(\sum_{i=0}^{t-1} a^i \epsilon_{t-i}, \sum_{i=0}^{t+h-1} a^i \epsilon_{t+h-i}\right) \\
&= \sigma_\epsilon^2 \left(a^h + a^{h+2} + ... + a^{h+2(t-1)}\right) \\
&= \begin{cases} \sigma_\epsilon^2 a^h \frac{1-a^{2t}}{1-a^2} & \text{if } |a| \neq 1 \\ \sigma_\epsilon^2 t & \text{if } |a| = 1. \end{cases}
\end{aligned}$$

Again we see the dependence on t, and again we see that a limit exists for $|a| < 1$ and is given by

$$\sigma_\epsilon^2 \frac{a^h}{1-a^2}.$$

This means that the series has an asymptotic stationary limit if $|a| < 1$. Thus we see that there exists a stationary AR(1) series (namely the limit of X_t above), and that this series has auto-covariance function

$$\gamma(h) = \sigma_\epsilon^2 \frac{a^{|h|}}{1-a^2}.$$

In particular, the variance of the process is

$$\sigma_X^2 = \mathbf{var}(X_t) = \sigma_\epsilon^2 \frac{1}{1-a^2}.$$

Thus the auto–correlation function is given by

$$\rho(h) = a^{|h|}.$$

If $|a| = 1$ we see that there do not exists a limit. If $|a| > 1$ the expression for the covariance diverges as well as $t \to \infty$. This could in principle be overcomed by the following trick.

The AR(1) model is given by the regression

$$X_t = aX_{t-1} + \epsilon_t,$$

which can also be written

$$X_{t-1} = \frac{1}{a} X_t - \frac{1}{a} \epsilon_t,$$

and in this setting the argument above could be carried through , but instead of representing X_t as a sum of white noise presently and previously recorded we would obtain a sum of future white noise values. This property is not desirable in time series analysis, since we have quite a clear definition of the time direction.

The process AR(1) with $|a| < 1$ is called causal or future independent. Non–causal processes shall hence not be considered in what follows.

Exercises

(1) Explore the graphs of the auto–correlation function of an AR(1) process for respectively positive and negative a, $|a| < 1$. Which connection is there between the graph of the auto–correlation function and the actual behaviour of the time–series itself? (hint: look at $a \to -1$.)

2.3 AR(2) processes

Auto–regressive processes of order two is the natural extension of the first–order model, to a linear regression on the previous two values. To be precise, an auto–regressive model of second order, AR(2), satisfies

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} = \epsilon_t, \tag{2.2}$$

where $\{\epsilon_t\}$ is a white noise, which we without loss of generality can assume to have zero mean and variance σ_ϵ^2.

Using the backward shift operator B, we can also write (2.2) as

$$(1 + a_1 B + a_2 B^2) X_t = \epsilon_t.$$

Factorizing the polynomial using its roots r_1, r_2,

$$\begin{aligned} 1 + a_1 B + a_2 B^2 &= (B - r_1)(B - r_2) \\ &= (r_1 - B)(r_2 - B) \\ &= r_1 r_2 (1 - \frac{1}{r_1} B)(1 - \frac{1}{r_2} B) \\ &= (1 - \frac{1}{r_1} B)(1 - \frac{1}{r_2} B) \\ &= (1 - \mu_1 B)(1 - \mu_2 B), \end{aligned}$$

since $r_1 r_2 = 1$ (put $B = 0$ in the polynomial and evaluate the factorization), and where $\mu_1 = 1/r_1$ and $\mu_2 = 1/r_2$. Note that since r_1 is a root of the polynomial $1 + a_1 z + a_2 z^2$ then $1 + a_1 r_1 + a_2 r_1^2 = 0$, and dividing by r_1^2 we get $\mu_1^2 + a_1 \mu_1 + a_2 = 0$, i.e. μ_1 is a root in the polynomial $z^2 + a_1 z + a_2$. Similarly, μ_2 is a root in the polynomial $z^2 + a_1 z + a_2$.

Using this factorization we get that

$$\begin{aligned} X_t &= \frac{\epsilon_t}{(1 - \mu_1 B)(1 - \mu_2 B)} \\ &= \frac{1}{\mu_1 - \mu_2} \left(\frac{\mu_1}{1 - \mu_1 B} - \frac{\mu_2}{1 - \mu_2 B} \right) \epsilon_t \end{aligned}$$

$$= \frac{1}{\mu_1 - \mu_2}\left(\sum_{s=0}^{\infty}\left(\mu_1^{s+1} - \mu_2^{s+1}\right)B^s\right)\epsilon_t$$
$$= \sum_{s=0}^{\infty}\frac{\mu_1^{s+1} - \mu_2^{s+1}}{\mu_1 - \mu_2}\epsilon_{t-s}.$$

To obtain the general solution to (2.2) we need to add to the expression above the solution to the equation

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} = 0.$$

This equation can be shown to have a solution on the form

$$c_1\mu_1^t + c_2\mu_2^t,$$

so the general solution to (2.2) is hence

$$c_1\mu_1^t + c_2\mu_2^t + \sum_{s=0}^{\infty}\frac{\mu_1^{s+1} - \mu_2^{s+1}}{\mu_1 - \mu_2}\epsilon_{t-s}.$$

For this expression to have a limit, as $t \to \infty$, we need at least $|\mu_1| < 1$ and $|\mu_2| < 1$ for the first term to vanish. But what about the second term? Under the condition that $|\mu_1| < 1$ and $|\mu_2| < 1$ we have that the variance of the second term is

$$\begin{aligned}\mathbf{var}\left(\sum_{s=0}^{\infty}\frac{\mu_1^{s+1} - \mu_2^{s+1}}{\mu_1 - \mu_2}\epsilon_{t-s}\right) &= \sum_{s=0}^{\infty}\left(\frac{\mu_1^{s+1} - \mu_2^{s+1}}{\mu_1 - \mu_2}\right)^2\sigma_\epsilon^2 \\ &< \infty.\end{aligned}$$

Thus the conditions $|\mu_1| < 1$ and $|\mu_2| < 1$ is enough to ensure asymptotic stationarity, and hence for a stationary AR(2) to exists.

We will not derive the auto–covariance function for this case, since the calculations will become rather tedious, and not provide any further intrisic inside into the structure of AR(2) processes.

What is interesting, however, is to consider the roots μ_1 and μ_2. They can both be complex valued or real valued, depending on whether the discriminat D of the second order equation $z^2 + a_1 z + a_2 = 0$ is negative or positive. In the former case the condition amounts to

$$D = a_1^2 - 4a_2 \leq 0 \implies a_2 \geq \frac{1}{4}a_1^2,$$

while in the latter we have

$$D = a_1^2 - 4a_2 > 0 \implies a_2 < \frac{1}{4}a_1^2.$$

In case of complex roots, we have that $D < 0$, and hence

$$\sqrt{a_1^2 - 4a_2} = i\sqrt{4a_2 - a_1^2}. \tag{2.3}$$

The solutions are respectively

$$\mu_1 = -\frac{a_1}{2} + \frac{\sqrt{a_1^2 - 4a_2}}{2}, \quad \mu_2 = -\frac{a_1}{2} - \frac{\sqrt{a_1^2 - 4a_2}}{2}.$$

Then using (2.3) we obtain

$$|\mu_1|^2 = \mu_1\bar{\mu}_1 = a_2, \quad |\mu_2|^2 = \mu_2\bar{\mu}_2 = a_2.$$

This means that

$$|\mu_1| = |\mu_2| = \sqrt{a_2}.$$

(note: $a_2 \geq 0$ is ensured by $D < 0$.) Then the stationarity condition becomes equivalent to $a_2 < 1$. The same results hold for $D = 0$, where we have two equal roots.

Let us consider the case of (two unequal) real roots. Since the roots are real they are the intersection between the graph of $f(z) = z^2 + a_1 z + a_2$ with the first axis. The condition $|\mu_1| < 1, |\mu_2| < 1$ then amounts to the insecting points to lie between -1 and 1. The minimum is attained for $-a_1/2$ and a sufficient condition to ensure the points to lie between -1 and 1 is hence

$$-1 \leq -\frac{a_1}{2} \leq 1, \quad f(1) > 0, f(-1) > 0.$$

This is then equivalent to

$$|a_1| \leq 2, \quad 1 + a_1 + a_2 \geq 0, \quad 1 - a_1 + a_2 \geq 0.$$

Noting that $|a_1| \leq 2 \implies a_2 \leq 1$ we have that for stationarity to occur in the limit, (a_1, a_2) must lie in the triangle defined by

$$a_2 \leq 1, \; a_1 + a_2 \geq -1, \quad a_1 - a_2 \leq 1.$$

In general, one can show that the variance of the AR(2) process is given by

$$\sigma_X^2 = \mathbf{var}(X_t) = \sigma_\epsilon^2 \frac{1 + a_2}{(1 - a_2)(1 - a_1 + a_2)(1 + a_1 + a_2)},$$

and the auto–correlation function,

$$\rho(h) = \frac{(1 - \mu_2^2)\mu_1^{h+1} - (1 - \mu_1^2)\mu_2^{h+1}}{(\mu_1 - \mu_2)(1 + \mu_1\mu_2)}, \quad h \geq 0.$$

This auto–correlation function behaves very differently for the cases of real or complex solutions.

Exercises
(1) Using the calculations above, show that the roots μ_1 and μ_2 in the complex case satisfies the following:
(i) μ_1 and μ_2 are complex conjugate, and can be written as

$$\mu_1 = \sqrt{a_2}e^{i\theta}, \quad \mu_2 = \sqrt{a_2}e^{-i\theta}.$$

(ii) The parameter θ is given by

$$\cos(\theta) = -\frac{a_1}{2\sqrt{a_2}}.$$

(iii) The auto–correlation function is given by

$$\rho(h) = \frac{a_2^{h/2}\left(\sin\left((h+1)\theta\right) - a_2\sin\left((h-1)\theta\right)\right)}{(1+a_2)\sin\theta}.$$

(2) Explore the different behavior of $\rho(h)$ for different values of μ_1 and μ_2 (distinguishing particularly between the real and the complex cases) plotting $\rho(h)$ in a computer program.

(3) For complex values of μ_1 and μ_2 we see some periodic behaviour of $\rho(h)$, that is, however, damping out as $h \to \infty$. Explain what effect this periodic behviour will have on the actual data. (hint: let $a_2 \to 1$, and show that $\rho(h) \to \cos\theta h$. Thus correlation $+1$ occur at multiples of $2\pi/\theta$. Conclude from this, that if $\rho(h)$ has an exact cyclic behavior, then so has X_t itself. Conclude the same for the damped periodic behavior.)

2.4 AR(p) and MA(p) processes

AR(p) processes are the natural generalization of AR(2) and AR(1) processes. A time series $\{X_t\}$ is called a auto–regressive process of order p if it satisfies

$$X_t + a_1X_{t-1} + a_2X_{t-2} + ... + a_pX_{t-p} = \epsilon_t,$$

where $\{\epsilon_t\}$ is a white noise process, and $a_1, ..., a_p$ are constants. Introducing the polynomial $\alpha(z) = 1 + a_1z + ... + a_pz^p$ we can in short form write

$$\alpha(B)X_t = \epsilon_t.$$

The auto–covariance function is less straightforward to calculate than for AR(1) and AR(2) processes, and we will omit the details.

Now we introduce another important concept in the model building of time series, the moving average processes. A moving average process of order p, $\{X_t\}$, is defined by

$$X_t = b_0\epsilon_t + b_1\epsilon_{t-1} + ... + b_p\epsilon_{t-d}$$

where $b_0, b_1, ..., b_p$ are constants and $\{\epsilon_t\}$ is a white noise. Defining $\beta(z) = b_0 + b_1 z + ... + b_p z^p$ we can write

$$X_t = \beta(B)\epsilon_t.$$

One of the main differences between auto–regressive processes and moving average processes is the following. In the auto–regressive case we have that X_t is a regression on its own past values plus a random error ϵ_t. Thus ϵ_t is going to influate on all future values $X_{t+1}, X_{t+2}, ...$ etc.. But in the moving average case, the influence of ϵ_t is no longer present after p future steps. This is the same as to say that the auto–correlation function for MA(p) processes will be zero after lag p, i.e.

$$\rho(h) = 0 \quad \text{for} \quad |h| > p.$$

The basic caracteristics, mean, variance and auto–correlation function are easy to calculate moving average processes, since they are composed of sums of independent random variables.

Starting with the variance, it is easy to see that

$$\sigma_X^2 = \mathbf{var}(X_t) = \sigma_\epsilon^2 \sum_{i=0}^{p} b_i^2,$$

where $\sigma_\epsilon^2 = \mathbf{var}(\epsilon_t)$. For $h \geq 0$ the covariance function is clearly given by

$$\mathbf{Cov}(X_t, X_{t+h}) = \begin{cases} \sigma_\epsilon^2 (b_0 b_h + b_1 b_{h+1} + ... + b_{p-h} b_p) & \text{if } 0 \leq h \leq p \\ 0 & \text{otherwise} \end{cases}.$$

Moreover the auto–correlation function satisfies

$$\rho(h) = \begin{cases} \sum_{i=h}^{p} b_i b_{i-h} / \sum_{i=0}^{p} b_i^2 & 0 \leq h \leq p \\ 0 & \text{otherwise} \end{cases}.$$

For both the auto–covariance function and auto–correlation function is should be noted that they are symmetrical, so for $h < 0$ we e.g. have that $\rho(h) = \rho(-h)$.

2.5 ARMA(p,q) processes

The class of ARMA processes, which are combined auto–regressive and moving average processes, constitute a flexible class of processes with the nice feature, that any stationary processes with $\lim_{n\to\infty} \gamma(h) = 0$ can be approximated arbitrarily close by an ARMA(p,q) processes. To this end, however, we may need p and q to be quite large, and that is in fact not desirable.

The more variables we include, the more we exhaust our data set, and the less data per parameter we are provided with. This means, as the number of variables grow, also the precision of the estimates become less precise. Later, in the section on estimation of ARMA processes, we shall present an approach by Akaike on automatic model selection. In his approach there is put a cost on each additional variable, and the estimation then becomes a matter of minimizing the cost: the model that is less costly is the preferred one.

The generality of the ARMA model is closely connected to that of rational functions (a rational function is a function that can be written on the form $p(x)/q(x)$, where $p(x)$ and $q(x)$ are polynomials). Any continuous function can be approximated arbitrarily close by rational functions.

Now turning to ARMA processes, an ARMA(p,q) process is a time series $\{X_t\}$ that satisfies

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} = b_0 \epsilon_t + b_1 \epsilon_{t-1} + ... + b_p \epsilon_{t-p} \tag{2.4}$$

Using the polynomial introduced in the previous section, we can write the condition (2.4) on the closed form

$$\alpha(B) X_t = \beta(B) \epsilon_t.$$

The intuitive meaning of this class of processes is that the regression of X_t on its past values does not only depend on the sampling error of X_t, ϵ_t, but as well on the sampling errors on the q previous variables. Here q can both be larger or smaller than p.

If p is larger then q the process is easy to interpret, as we consider a regression on previous values, where some of the variables error terms are taken into account. If on the other hand q is larger than p is may be difficult to give any justifucation for such a model other than it fits the data well.

2.6 The partial auto–correlation function

For a time series $X_1, X_2, ..., X_t, ...$, consider for a given k the regression of X_{k+1} and X_1 on their intermediate values $X_2, ..., X_k$, i.e. the model

$$X_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + ... + \alpha_k X_k$$

$$X_{k+1} = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k.$$

In these linear regressions we then estimate (least squares) the parameters, obtaining the corresponding estimates $\hat{\alpha}_1, ..., \hat{\alpha}_k, \hat{\beta}_1, ..., \hat{\beta}_k$. Then form the residuals

$$\begin{aligned} R_1 &= X_1 - (\hat{\alpha}_1 + \hat{\alpha}_2 X_2 + ... + \hat{\alpha}_k X_k) \\ R_{k+1} &= X_{k+1} - \left(\hat{\beta}_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k\right). \end{aligned}$$

The partial auto-correlation function $\alpha(h)$ is then defined as follows:

$$\begin{aligned} \alpha(1) &= \mathbf{corr}(X_2, X_1) \\ \alpha(k) &= \mathbf{corr}(R_{k+1}, R_1). \end{aligned}$$

So what is the interpretation of the partial auto-correlation function? It is somehow the correlation between X_1 and X_{k+1} after taking into account the intermediate steps. By regressing X_1 on $X_2, ..., X_k$ we try to express X_1 in terms of $X_2, ..., X_k$. Similarly we try to express X_{k+1} in terms of $X_2, ..., X_k$. The deviation from this exact linear relationship, the residuals in other words, then tells us how closely related are X_1 and X_{k+1}.

For an AR(p) process it is clear that $\alpha(h) = 0$ for $h > p$. This follows from the fact that in the AR(p) process $X_{p+2} = c_0 + c_{p+1}X_{p+1} + ... + c_2 X_2 + \epsilon_{p+2}$, so the residual is ϵ_{p+2}, which is independent of $X_1, ..., X_{p+1}$, and hence also independent of $X_1 - d_1 - d_2 X_2 - ... - d_{p+1}X_{p+1}$, which immediately implies that $\alpha(h) = 0$ for $h > p$.

This is not the case for MA(q) processes. For example the MA(1) process has partial auto-correlation function

$$\alpha(h) = \frac{(-b)^h(1 - b^2)}{1 - b^{2(k+1)}},$$

which will not vanish after a certain point.

How to calculate the partial auto-correlation functions in practice for concrete ARMA processes involves the calculations of the auto-correlations as well, and we will therefore omit any further excursions in that direction. In practice, however, it provides valuable information when checking if a model is an AR process.

2.7 Generalized linear processes

By a generalized linear process we understand a time series $\{X_t\}$ that can be represented as

$$X_t = \sum_{u=0}^{\infty} g_u \epsilon_{t-u},$$

where the g_u's are constants such that $\sum_{u=0}^{\infty} g_u^2 < \infty$ and $\{\epsilon_t\}$ is a white noise process. It follows directly from the definition that we may consider the generalized linear process as an infinite moving average process, MA(∞).

The variance of this process is readily calculated to be

$$\begin{aligned}
\sigma_X^2 &= \mathbf{var}(X_t) \\
&= \mathbf{var}\left(\sum_{u=0}^{\infty} g_u \epsilon_{t-u}\right) \\
&= \sum_{u=0}^{\infty} g_u^2 \sigma_\epsilon^2 \\
&= \left(\sum_{u=0}^{\infty} g_u^2\right) \sigma_\epsilon^2,
\end{aligned}$$

where σ_ϵ^2 denotes the variance of the white noise process, $\sigma_\epsilon^2 = \mathbf{var}(\epsilon_t)$. Thus the generalized linear process has finite variance.

Put $g_u = 0$ for $u < 0$. Then we can write $X_t = \sum_{u=-\infty}^{\infty} g_u \epsilon_{t-u}$. Then the auto-covariance function is

$$\begin{aligned}
\gamma(h) &= \mathbf{Cov}(X_t, X_{t+h}) \\
&= \mathbf{Cov}(X_t, X_{t-h}) \\
&= \mathbf{Cov}\left(\sum_{u=-\infty}^{\infty} g_u \epsilon_{t-u}, \sum_{u=-\infty}^{\infty} g_u \epsilon_{t-h-u}\right) \\
&= \sigma_\epsilon^2 \sum_{u=-\infty}^{\infty} g_u g_{u-h}. \qquad (2.5)
\end{aligned}$$

The convergence of the series (2.5) is ensured by

$$|\mathbf{Cov}(X_t, X_{t-h})| \leq \sqrt{\mathbf{var} X_t}\sqrt{\mathbf{var} X_{t-h}} = \sigma_X^2 < \infty.$$

The auto-correlation function is given by

$$\rho(h) = \gamma(h)/\sigma_\epsilon^2 = \sum_{u=-\infty}^{\infty} g_u g_{u-h} \Big/ \sum_{u=0}^{\infty} g_u^2 .$$

Thus the condition $\sum_{u=-\infty}^{\infty} g_u^2 < \infty$ ensures second order stationarity as well.

Consider the series

$$G(z) = \sum_{u=0}^{\infty} g_u z^u.$$

This series (which is defined in terms of the complex variable z) exists (converges) if e.g.

$$\sum_{u=0}^{\infty} |g_u||z|^u < \infty.$$

This means for $|z| \leq 1$ the condition $\sum_{u=1}^{\infty} |g_u| < \infty$ is sufficient. Imposing this (stronger) condition on the series g_u, we then have that $G(z)$ exists for $z \leq 1$.

The existence of $G(z)$ for $|z| \leq 1$ is the same as to say that $G(z)$ is analytic inside and on the unit cirle. The importance of the function $G(z)$ is from the relationship

$$X_t = G(B)\epsilon_t.$$

If we moreover assume that $G^{-1}(z) < \infty$ for $|z| \leq 1$ then $G^{-1}(z)$ is also analytic inside and on the unit circle, and hence it allows for a power series expansion

$$G^{-1}(z) = \sum_{u=0}^{\infty} h_u z^u,$$

where $\sum_{u=0}^{\infty} |h_u| < \infty$. Thus we can write the infinite moving average process as an infinite order auto–regressive process as well, namely

$$G^{-1}(z)X_t = \epsilon_t.$$

Looking at the ARMA process

$$\alpha(B)X_t = \beta(B)\epsilon_t,$$

the condition for X_t to be a general linear process is that $G(z) = \beta(z)/\alpha(z)$ is analytic inside and on the unit cirle, and since $\beta(z)$ is a polynomial and hence analytical, the condition is only imposed on $\alpha(z)$, namely that $\alpha(z)^{-1}$ analytic inside and on the unit circle, which again equivalent to $\alpha(z)$ does not have any roots inside or on the unit circle.

On the other hand, by a symmetric argument, the condition for the ARMA process being expressible as an (possibly infinite) auto–regressive model, is that $\beta(z)$ does not have any roots on or inside the unit cirle.

If a process can be expressed as a generalized linear model it is called causal (or future independent), and if can be expressed as an infinite order auto–regressive model, it is called invertible.

It is the simple form of the auto–covariance function (2.5) that forms the basis for the calculations of auto–covariance functions of ARMA processes in practice. All we need to do is to represent the (causal) ARMA model as a generalized linear model (finding the g_u's) and calculate (2.5).

Consider the usual ARMA(p,q) model,

$$\alpha(B)X_t = \beta(B)\epsilon_t.$$

Assuming the model is causal,

$$G(z) = \frac{\beta(z)}{\alpha(z)}$$

is analytical for $|z| \leq 1$, and hence (by definition of analytic) is has a power series expansion

$$G(z) = \sum_{u=0}^{\infty} \xi_u z^u.$$

The ξ_u's are necessarily the coefficients g_u we are looking for, i.e. $g_u = \xi_u$. So in principle we could differenciate $G(z)$ u times and evaluate at 0 to find g_u, but there exists in fact easier methods.

One of them is to rewrite $G(z) = \beta(z)/\alpha(z)$ as $G(z)\alpha(z) = \beta(z)$ and calculate the coefficient of z^j directly from this expression. Take for example the causal ARMA(2,1) model,

$$(1 - B + \frac{1}{4}B^2)X_t = (1 + B)\epsilon_t.$$

Then $\alpha(z) = 1 - z + \frac{1}{4}z^2$ and $\beta(z) = (1 + z)$. Then

$$\begin{aligned} G(z)\alpha(z) &= \left(\sum_{u=0}^{\infty} g_u z^u\right)\left(1 - z + \frac{1}{4}z^2\right) \\ &= \sum_{u=0}^{\infty} g_u z^u + \sum_{u=0}^{\infty}(-1)g_u z^{u+1} + \sum_{u=0}^{\infty}\frac{1}{4}g_u z^{u+2}, \end{aligned}$$

and then equate from

$$\sum_{u=0}^{\infty} g_u z^u + \sum_{u=0}^{\infty}(-1)g_u z^{u+1} + \sum_{u=0}^{\infty}\frac{1}{4}g_u z^{u+2} = 1 + z.$$

On the left hand side (LHS) there is only on coefficient to z^0 namely g_0 in the first sum. Comparing this to the RHS we get $g_0 = 1$. There are two coefficients to z, from the first sum on the RHS we get g_1 (for $u = 1$) and from the second sum $(-1)g_0$ (for $u = 0$). Comparing to the RHS we have the equation

$$g_1 - g_0 = 1,$$

and since $g_0 = 1$ then $g_1 = 2$. For coefficients to z^2 we get the equation

$$g_2 - g_1 + \frac{1}{4}g_0 = 0,$$

which gives $g_2 = \frac{7}{4}$. In general, for $u \geq 2$, we have that

$$g_u - g_{u-1} + \frac{1}{4} g_{u-2} = 0$$

the solution of which is

$$g_u = (a + nb)2^{-n}.$$

The constants are then verified to be $a = 1$ and $b = 1$ using $g_0 = 1$ and $g_1 = 2$.

This method is cumbersome in practice, but still bar far easier than differentation.

Chapter 3

Estimation and Order Selection

3.1 Introduction

In this chapter we introduce the most important concepts from estimation theory and order selection methods for ARMA processes. Though we omit most of the proofs it is the intention that the exposition should provide a good feeling for the practical and conceptional problems involved in any data study of such series.

Modern estimation procedures and order selection techniques are presented, and particular details in the analysis phase is highlighted as well.

3.2 Estimation of mean.

Consider a (second order) stationary time series, $\{X_t\}$, $t \in \mathbb{Z}$, and let

$$\begin{aligned} \mathbb{E}(X_t) &= \mu \\ \mathbf{var}(X_t) &= \sigma^2 \\ \mathbf{Cov}(X_t, X_{t+h}) &= \gamma(h) = \sigma^2 \rho(h). \end{aligned}$$

If we are confronted with a data set $x_1, ..., x_N$ of data from $\{X_t\}$, in general all parameters are unknown including the form of the auto–covariance function. All these parameters have to be estimated from data.

Standard statistical procedures, like the maximum likelihood approach, require full knowledge about the distribution of $(X_1, ..., X_N)$ in order to maximize the likelihood function over the data.

In time series analysis we only assume second order stationarity, which in fact only put restrictions on the first two moments. This means that we do not have any knowledge about the distributions of $X_1, ..., X_N$, and even worse they are allowed to be differently

distributed. Another feature that complicates the matter is the assumption of correlation between variables. Indeed in the likelihood approach it is standard to assume i.i.d. variables, which simplifies the likelihood function considerably.

After these comments the reader should understand the reason for why it has not been custom to apply more established statistical procedures to time series, but instead with some intuitive approaches, which some critics instead may call ad hoc. These approaches still play an important role though maximum likelihood estimation has become standard even in cases where the data show a clear deviation from normality.

Using the fact that all X_i have the same mean, μ, suggests that the sample mean $\bar{X}_N$, should be an useful estimator of the mean,

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Note that any estimator is a random variable (hence written with capital letters) per definition, and we can therefore explore their distributional and sampling properties in terms of the original variables $X_1, ..., X_N$. When estimating the sample mean, however, we will write $\bar{x}_N$ as the corresponding realization of $\bar{X}_N$.

The sample mean is an unbiased estimator of μ, since

$$\mathbb{E}\bar{X}_N = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}X_i = \mu.$$

The variance of $\bar{X}$ is

$$\begin{aligned}
\mathbf{var}(\bar{X}_N) &= \mathbf{var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \\
&= \frac{1}{N^2}\mathbf{var}\left(\sum_{i=1}^{N} X_i\right) \\
&= \frac{1}{N^2}\mathbf{Cov}\left(\sum_{i=1}^{N} X_i, \sum_{i=1}^{N} X_i\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{Cov}(X_i, X_j) \\
&= \frac{\sigma^2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\rho(j-i) \\
&= \frac{\sigma^2}{N^2}\sum_{i=-(N-1)}^{N-1}(N-|i|)\rho(i)
\end{aligned}$$

$$= \frac{\sigma^2}{N} \sum_{i=-(N-1)}^{N-1} (1 - \frac{|i|}{N})\rho(i).$$

Taking absolut values we obtain

$$\left|N\mathbf{var}(\bar{X}_N)\right| \leq \sigma^2 \sum_{|h|<N} |\rho(h)|.$$

If $\rho(h) \to 0$ as $h \to \infty$ we have

$$\frac{1}{N} \sum_{|h|<N} |\rho(h)| \to 0$$

as $N \to \infty$. To see this, let $m < n$ and write

$$\begin{aligned}
\frac{1}{n} \sum_{|h|<n} |\rho(h)| &= \frac{1}{n} \sum_{|h|<m} |\rho(h)| + \frac{1}{n} \sum_{m \leq |h|<n} |\rho(h)| \\
&\leq \frac{1}{n} \sum_{|h|<m} |\rho(h)| + 2\frac{n-m}{n} \max\{|\rho(m)|, |\rho(m+1)|, ..., |\rho(n-1)|\} \\
&= \frac{1}{n} \sum_{|h|<m} |\rho(h)| + 2\left(1 - \frac{m}{n}\right) \max\{|\rho(m)|, |\rho(m+1)|, ..., |\rho(n-1)|\} \\
&\to 2\max\{|\rho(m)|, |\rho(m+1)|, ..\}
\end{aligned}$$

as $n \to \infty$. Then for all m we have that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{|h|<n} |\rho(h)| \leq 2\max\{|\rho(m)|, |\rho(m+1)|, ...\}$$

and letting $m \to \infty$ then gives that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{|h|<n} |\rho(h)| = 0.$$

But then we have proved that as $n \to \infty$,

$$\mathbf{var}(\bar{X}_n) \to 0,$$

if $\rho(h) \to 0$ as $h \to \infty$. Moreover, if $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$ then

$$\begin{aligned}
\lim_{n\to\infty} n\mathbf{var}(\bar{X}_n) &= \lim_{n\to\infty} \sum_{|h|<n} \left(1 - \frac{|h|}{n}\right) \gamma(h) \\
&= \sum_{h=-\infty}^{\infty} \gamma(h).
\end{aligned}$$

The condition $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$ is satisfied by all ARMA processes. So for such processes with $\nu = \sum_{h=-\infty}^{\infty} \gamma(h) \neq 0$ we have that

$$\bar{X}_n \sim N(\mu, \nu/n). \tag{3.1}$$

If the data are normally distributed, we have the exact property

$$n^{1/2}\left(\bar{X}_n - \mu\right) \sim N\left(0, \sum_{|h|<n}\left(1 - \frac{|h|}{n}\right)\gamma(h)\right). \tag{3.2}$$

These results are very useful for finding large–sample confidence or approximate confidence intervals for μ.

Example 3.2.1 100 values where simulated of an ARMA(2,2) process with mean 0 and variance 1. The emperical mean was calculated to be $\bar{x} = 0.059582$. From the estimated auto–covariance function we calculated

$$\sum_{h=-40}^{40} \hat{\gamma}(h) = 4.929905.$$

Note we only used 40 values from the auto–covariance function. The reason is simply that the time series package applied, PEST, does not provide more data than that, and moreover that we can never use the theoretical full amount 100 since the estimate will be very poor for larger lags; we can only estimate the auto–correlation function up to about lag $N/4$, where N is the size of the data set. In practice this does not matter much since the auto–covariances have exponential tails, and hence the information we loose will be insignificant.

Then the 95 % confidence interval for μ based on the asymptotic normality (3.1) is given by

$$\bar{x} - 1.96\sqrt{\nu/n} \leq \mu \leq \bar{x} + 1.96\sqrt{\nu/n}$$

which numerically is equivalent to

$$-0.3756 \leq \mu \leq 0.4948.$$

If we actually knew that the ARMA(2,2) series was Gaussian (has a normal distribution) then we can apply the exact formula (3.2). In this case we have that

$$\sum_{h=-40}^{40} (1 - \frac{|h|}{n})\hat{\gamma}(h) = 5.463340$$

which results in the confidence interval

$$-0.3985 \leq \mu \leq 0.5177.$$

PEST provide the usual standard deviation of data, which in this case was 1.4977. We cannot construct the confidence interval using this standard deviation, since this would be the same as to assume independence, and the confidence interval we obtain in this case is only

$$-0.1803 \leq \mu \leq 0.2994.$$

This may seem as a stronger statement than the other two confidence intervals, but that this confidence interval is much tighter than the other two simply means that the tests at the same levels (e.g. 99 %) are more prawn to reject the hypothesis $\mu = 0$ using the independence assumption than using the exact or asymptotic formula under the stationarity assumption.

Note that PEST does not automatically provide confidence intervals for the mean using the stationarity assumption. But what we can do is to save the auto–covariances in a file, and use a little program (or a calculator!) to calculate the variances above. In this example we used FORTRAN programs. In the case of exact variance it looks like:

```
OPEN(UNIT=8,FILE='arma22.acf',STATUS='old')
SUM=0
DO I=1,40
READ(8,*) X1
X=1.0*T
T=1-X/100
SUM=SUM+T*X1
ENDDO
V=2*SUM+1.4977*1.4977
WRITE(*,*) V
END
```

It should be underlined already at this point, that when a statistical analysis is presented all estimators should be available together with their confidence intervals, and, if available, their exact or asymptotic distributions. A report that does not specify confidence intervals for the estimators cannot be trusted in its conclusions.

Exercise: Simulate some stationary (ARMA) processes in PEST, and calculate confidence intervals for the mean values. Do the same for real data that has been put into stationary mode. Use e.g. the data 'elec'.

3.3 Estimation of the auto–covariance function

The estimation theory for auto–covariances and auto-correlations are by far more complicated than for the mean, and it is beyond the scope of these notes to outline any details for the distributional properties of the estimates.

As estimator for the auto–covariance function we will use

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n) \quad 0 \leq h \leq n-1, \tag{3.3}$$

and for the auto–correlation

$$\hat{\rho}(h) = \frac{\hat{\gamma(h)}}{\hat{\gamma(0)}}. \tag{3.4}$$

Note that for large h (like $h = n-1$) (3.3) provides a very poor estimate for the auto–covariance function, since it is based on very few values. In practice one should not rely on estimates beyond lag $N/4$, where N is the number of data. The experienced reader may ask why we don't use $1/(n-h)$ as normalizing factor in the average instead of $1/n$ to obtain a 'proper' average. The reason is purely technical, and has to do with the sampling properties; it has been claimed to have a smaller quadratic error than the other, and it has some desirable properties in terms of matrices (positive definitness).

We shall not go into further derivations in this section, but merely discuss the important results.

Consider the general linear process (now not necessarily causal)

$$X_t = \mu + \sum_{j=-\infty}^{\infty} g_j \epsilon_j,$$

where $\{\epsilon_t\}$ are i.i.d's (and not only a white noise process). Fix a lag h, and let W be the matrix who's ij element $(i,j = 1,...h)$ is given by

$$\begin{aligned} w_{ij} &= \sum_{k=1}^{\infty}\{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \\ &\quad \times \{\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\}. \end{aligned}$$

Then one can show that if $\mathbb{E}\epsilon_t^4 < \infty$ and $\sum_{j=-\infty}^{\infty}|g_j| < \infty$ then asymptotically,

$$\hat{\boldsymbol{\rho}}(h) \sim N(\boldsymbol{\rho}(h), n^{-1}W),$$

where

$$\hat{\boldsymbol{\rho}}(h) = \begin{pmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \\ . \\ . \\ \hat{\rho}(h) \end{pmatrix}, \quad \boldsymbol{\rho}(h) = \begin{pmatrix} \rho(1) \\ \rho(2) \\ . \\ . \\ \rho(h) \end{pmatrix}.$$

We can also remove the restriction on the error term, but then we have to impose the following extra condition

$$\sum_{j=-\infty}^{\infty} |j| g_j^2 < \infty.$$

These conditions are satisfied by any ARMA process, where the white noise process is not only uncorrelated but actually independent. In practice this is not a severe restriction, and certainly not if we are dealing with Gaussian processes, since we then have equivalence between uncorrelatedness and independence.

Again, in presenting the final report on a time series analysis, it would be professional to present the covariance matrix of the auto–correlations. To calculate W in practice we simply replace $\rho(k)$ by $\hat{\rho}(k)$.

If X_t is an i.i.d. sequence then $W = I$ (the identity matrix, $w_{ij} = 1$ if $i = j$ and 0 otherwise), and we have that asymptotically

$$\hat{\boldsymbol{\rho}}(h) \sim N(\hat{\boldsymbol{\rho}}, n^{-1} I).$$

This means in particular that

$$\mathbf{var}(\hat{\rho}(h)) \sim \frac{1}{n}$$

and hence the 95 % confidence interval for all $\hat{\rho}(h)$'s is

$$\hat{\rho(h)} - \frac{1.96}{\sqrt{n}} \leq \rho(h) \leq \hat{\rho(h)} + \frac{1.96}{\sqrt{n}}.$$

So if we draw the 95 % lines into the graph of $\hat{\rho(h)}$, then all points $\hat{\rho(h)}, h \neq 0$ should be between these two lines if X_t is i.i.d. Often we will use this fact in selecting models. In particular to identify i.i.d. processes.

3.4 Estimation of parameters in AR(p) processes: Yule–Walker equations

Throughout we will assume that the mean has been subtracted from the process, so that we are considering a zero mean AR(p) process on the form

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} = \epsilon_t, \tag{3.5}$$

where then ϵ_t is a white noise process with zero mean. Suppose moreover that the process we are considering is causal. Multiply (3.5) with X_{t-k}, $k \geq 0$ and obtain the equation

$$X_t X_{t-k} + a_1 X_{t-1} X_{t-k} + a_2 X_{t-2} X_{t-k} + ... + a_3 X_{t-p} X_{t-k} = \epsilon_t X_{t-k}. \tag{3.6}$$

Take expectation to get

$$\begin{aligned} \mathbb{E}(X_t X_{t-k}) + a_1 \mathbb{E}(X_{t-1} X_{t-k}) + a_2 \mathbb{E}(X_{t-2} X_{t-k}) + \ldots &+ a_p \mathbb{E}(X_{t-p} X_{t-k}) \\ &= \mathbb{E}(\epsilon_t X_{t-k}). \end{aligned} \tag{3.7}$$

We have assumed that the process is causal, so we can write

$$X_t = \sum_{u=0}^{\infty} g_u \epsilon_{t-u}$$

and hence for $k > 0$

$$\mathbb{E}(\epsilon_t X_{t-k}) = \mathbf{Cov}\left(\sum_{u=0}^{\infty} g_u \epsilon_{t-k-u}, \epsilon_t \right) = 0.$$

Thus we have the equation

$$\gamma(k) + a_1 \gamma(k-1) + \ldots + a_p \gamma(k-p) = 0, \tag{3.8}$$

for $k = 1, \ldots, p$, which in the literature is known as the Yule–Walker equation.

If we want to estimate the coefficients $a_1, \ldots, a_p$ we can also make use of the Yule–Walker equation, replacing $\gamma(h)$ by its estimate $\hat{\gamma}(h)$ and solve for $a_1, \ldots, a_p$. To this end we establish p linear equations as follows:

$$\begin{aligned} \hat{\gamma}(1) &+ \hat{a}_1 \hat{\gamma}(0) + \hat{a}_2 \hat{\gamma}(-1) + \ldots + \hat{a}_p \hat{\gamma}(1-p) = 0 \\ \hat{\gamma}(2) &+ \hat{a}_1 \hat{\gamma}(1) + \hat{a}_2 \hat{\gamma}(0) + \ldots + \hat{a}_p \hat{\gamma}(2-p) = 0 \\ \hat{\gamma}(3) &+ \hat{a}_1 \hat{\gamma}(2) + \hat{a}_2 \hat{\gamma}(1) + \ldots + \hat{a}_p \hat{\gamma}(3-p) = 0 \\ &\ldots. \\ \hat{\gamma}(p) &+ \hat{a}_1 \hat{\gamma}(p-1) + \hat{a}_2 \hat{\gamma}(p-2) + \ldots + \hat{a}_p \hat{\gamma}(0) = 0. \end{aligned}$$

Write

$$\hat{\boldsymbol{a}} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ . \\ . \\ \hat{a}_p \end{pmatrix}, \hat{\boldsymbol{\gamma}}_p = \begin{pmatrix} \hat{\gamma}(1) \\ \hat{\gamma}(2) \\ . \\ . \\ \hat{\gamma}(p) \end{pmatrix}$$

and

$$\hat{\boldsymbol{\Gamma}}_p = \begin{pmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \hat{\gamma}(2) & \ldots & \hat{\gamma}(p-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \hat{\gamma}(1) & \ldots & \hat{\gamma}(p-2) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \hat{\gamma}(p-1) & \hat{\gamma}(p-2) & \hat{\gamma}(p-3) & \ldots & \hat{\gamma}(0) \end{pmatrix}.$$

Then we can write the p equations on compact form as

$$\hat{\Gamma}_p \hat{\boldsymbol{a}} = -\hat{\gamma}_p. \tag{3.9}$$

The mean has been subtracted, so all we need to estimate now is the variance of the white noise process, $\hat{\sigma}_\epsilon^2$. Using (3.7) with $k = 0$ gives after taking expectations

$$\gamma(0) + a_1\gamma(1) + ... + a_p\gamma_p = \sigma_\epsilon^2,$$

so we estimate the variance by

$$\hat{\gamma}(0) + \hat{a}_1\hat{\gamma}(1) + ... + \hat{a}_p\hat{\gamma}(p) = \hat{\sigma}_\epsilon^2,$$

or written on compact form

$$\hat{\sigma}_\epsilon^2 = \hat{\gamma}(0) + \hat{\boldsymbol{a}}'\hat{\gamma}.$$

The Yule–Walker estimates $\hat{a}_1, ..., \hat{a}_p$ have the following asymptotic limit: If the white noise process is in fact an i.i.d. sequence, then

$$n^{1/2}(\hat{\boldsymbol{a}} - \boldsymbol{a}) \Longrightarrow N(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{\Gamma}_\mathbf{p}^{-1}),$$

where $\Longrightarrow$ means convergence in distribution, $\boldsymbol{a} = (a_1, ..., a_p)'$, $\mathbf{0}$ the vector of zeroes and $\boldsymbol{\Gamma}_p = \{\gamma(i-j)\}_{i,j=1,...,p}$.

In practice to obtain confidence intervals for the estimates one would replace all parameters in the distributional limit with its corresponding estimates.

3.5 Other techniques for estimation of AR processes

The Yule–walker approach is in fact only one approach among many existing ones, and certainly not the most precise one. It is useful, however, because it is simple and fast to calculate, and does not impose additional assumptions on the distributions of the time series.

The following methods are all superior to the Yule–Walker approach, and are listed in decreasing order of precision, such that the first one is the most precise method.

3.5.1 Maximum likelihood method

This method is only feasible if the white noise process is normally distributed, which in turn means that the whole time series itself is normally distributed. To write down the covariance matrix of the time series is a complicated matter, and we omit it. All we need to mention at this point is that the maximum likelihood estimate can be calculated using an iterative procedure, which is slower than the Yule–Walker approach. The program PEST, and most other packages, have the option of maximum likelihood, and is greatly recommended when the data can be assumed to be normally distributed. This assumption should of course be checked.

3.5.2 Least squares methods

If the data show a significant deviation from normality, it is recommendable not to use the maximum likelihood approach, but rather a least squares method, or approximate least squares method. The least squares method and the maximum likelihood method are equivalent when the data are Gaussian.

3.5.3 Durbin–Levinson algorithm

The Durbin–Levinson algorithm is a convenient way of estimating parameters in autoregressive processes, and is used by PEST. In particular when considering order selection of the AR process it is useful. Let us suppose that the AR model under consideration is of order p, but p is unknown. We then attempt to fit an AR model of order m with parameters $\hat{\phi}_{m1}, ..., \hat{\phi}_{mm}$. If $p < m$ then we would expect the parameter $\hat{\phi}_{mm}$ to be very small, so in an order selection procedure we would fit models for increasing m until $\hat{\phi}_{mm}$ becomes sufficiently small. Now the question is: what is sufficiently small.

It can be shown that $\hat{\phi}_{mm}$ is an estimate for the partial auto–correlation function at lag m, and we know that the partial auto–correlation function will vanish for lags larger than p if the model is auto–regressive of order p. The Durbin–Levinson algorithm calculates recursively the parameters $\hat{\phi}_{m1}, ..., \hat{\phi}_{mm}$ using a sort of best linear predictor method. The following result is most important when checking for the model to be an AR model of order p. If the white noise process is also i.i.d., then the estimated partial auto–correlations satisfy the following limit:

$$n^{1/2}\hat{\phi}_{mm} \Longrightarrow N(0,1),$$

as $n \to \infty$. This means as the sample size n becomes large, the estimated partial auto–correlation function will have 95 % of its function values larger than lag p placed in the interval $[-1.96n^{-1/2}, 1.96n^{-1/2}]$.

We now formulate the actual content of the Durbin–Levinson algorithm. If we have a model with $\hat{\gamma}(0) > 0$, then the fitted parameters in the auto–regressive model can be calculated by the following recursive scheme: Let $\hat{\boldsymbol{\phi}}_m = (\hat{\phi}_{m1}, ..., \hat{\phi}_{mm})'$ and let $\hat{v}_j$ be the estimated white noise variance in the AR(j) model.

$$\begin{aligned}
\hat{\phi}_{11} &= \hat{\rho}(1) \\
\hat{v}_1 &= \hat{\gamma}(0)\left(1 - \hat{\rho}(1)^2\right) \\
\hat{\phi}_{mm} &= \frac{1}{\hat{v}_{m-1}}\left(\hat{\gamma}(m) - \sum_{j=1}^{m-1} \hat{\phi}_{m-1,j}\hat{\gamma}(m-j)\right)
\end{aligned}$$

$$\begin{pmatrix} \hat{\phi}_{m1} \\ \hat{\phi}_{m2} \\ .. \\ \hat{\phi}_{m,m-1} \end{pmatrix} = \hat{\boldsymbol{\phi}}_{m-1} - \hat{\phi}_{mm} \begin{pmatrix} \hat{\phi}_{m-1,m-1} \\ \hat{\phi}_{m-1,m-2} \\ .. \\ \hat{\phi}_{m-1,1} \end{pmatrix}$$

$$\hat{v}_m = \hat{v}_{m-1}(1 - \hat{\phi}_{mm}^2).$$

This scheme calculates estimated parameters of the model of order $m+1$ using the fitted parameters of the model of order m.

To calculate the confidence intervals, let $\hat{\boldsymbol{\Gamma}}_p$ be the matrix $\{\hat{\gamma}(i-j)\}_{i,j=1,\ldots,p}$ and let $\hat{v}_{ij}$ be the ij th element of $\hat{v}_p \hat{\boldsymbol{\Gamma}}_p^{-1}$. Then the confidence interval of level $1-\alpha$ (e.g. $\alpha = 0.05$ for 95 % confidence limits) for ϕ_{pj} is given by

$$\hat{\phi}_{pj} - n^{-1/2}\Phi_{1-\alpha/2}\hat{v}_{jj}^{1/2} \leq \phi_{pj} \leq \hat{\phi}_{pj} + n^{-1/2}\Phi_{1-\alpha/2}\hat{v}_{jj}^{1/2}, \tag{3.10}$$

where $\Phi_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution (N(0,1)).

Here the matrix $\hat{v}_p\hat{\Gamma}_p^{-1}$ plays the role of estimated standard deviation of $\hat{\boldsymbol{\phi}}_p$, the diagonal elements of which v_{ii} hence are the estimated standard deviations of $\hat{\phi}_{pi}$. The program PEST does not calculate the confidence intervals for the parameters of the model, but provides the standard deviation of the estimators, and from those we can then construct the confidence intervals. For 95 % confidence intervals this is readily done by multiplying the standard deviation on the estimator by 1.96, cfr. the method above.

The Durbin–Levinson algorithm is not as precise as the maximum likelihhod or least squares methods, but it is much faster. The main use of Durbin–Levinsons algorithm is to create initial values for the maximum likelihood estimation.

The maximum likelihood and least squares methods are both iterative methods that in practice are rather slow. Therefore it is important when applying one of these methods to come up with a good initial guess, that is a guess that is reasonable close to the maximum likelihood or the quadratic minimum error. This is where Durbin–Levinson comes in, and provides very qualified initial values.

Exercise

For a concrete AR model (free choice) calculate estimates and their confidence intervals using the standard deviations on the parameters provided by PEST, and by applying Durbin–Levinson directly to the estimated auto–covariance function.

3.6 Estimation of parameters in MA processes: The innovation algorithm

The situation for MA processes is very similar to that of AR process, only with that important difference that the algorithm for calculating the estimates is different. It is

called the innovation algorithm and works as follows.

Consider the MA(p) process

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_p \epsilon_p,$$

where ϵ is a zero mean white noise. As in the case of the AR process we do not know the order of the moving average beforehand, so for each possible order m we may apply an estimation procedure to obtain estimates of the parameters $\hat{\theta}_{m1}, \hat{\theta}_{m2}, ..., \hat{\theta}_{mm}$ and of the white noise variance $\hat{v}_m$. A fast way of estimating the parameters i through the innovations algorithm: If the process is so that $\gamma(0) > 0$ then the recursion scheme is as follows

$$\begin{aligned}
\hat{v}_0 &= \hat{\gamma}(0) \\
\hat{\theta}_{m,m-k} &= \frac{1}{\hat{v}_k}\left(\hat{\gamma}(m-k) - \sum_{j=0}^{k-1} \hat{\theta}_{m,m-j}\hat{\theta}_{k,k-j}\hat{v}_j\right), \quad k = 0, ..., k-1 \\
\hat{v}_m &= \hat{\gamma}(0) - \sum_{j=0}^{m-1} \hat{\theta}^2_{m,m-j}\hat{v}_j.
\end{aligned}$$

The use of this algorithm is similar to that of Durbin–Levinsons algorithm, and the part we will be most interested in is to obtain confidence intervals for the estimated parameters. The 95% confidence interval for the parameter θ_{mj} is given by

$$\hat{\theta}_{mj} - 1.96n^{-1/2}\left(\sum_{k=0}^{j-1}\hat{\theta}^2_{mk}\right)^{1/2} \leq \theta_{mj} \leq \hat{\theta}_{mj} + 1.96n^{-1/2}\left(\sum_{k=0}^{j-1}\hat{\theta}^2_{mk}\right)^{1/2}.$$

The use of the innovation algorithm is like the Durbin–Levinson algorithm not the most efficient one in terms of precision of the estimates; to that end we still have to refer to maximum likelihood and least squares methods. But the innovations algorithm has the advantage that it is fast, and provides us with good starting values for the maximum likelihood or least squares iterative procedures, as was the case for AR processes using Durbin–Levinsons algorithm. Therefore the innovations algorithm is also considered as a preliminary estimation procedure, that serves as initial values for maximum likelihood or least squares methods.

Finally we comment that the preliminary estimation of causal ARMA processes uses the innovations algorithm as well, as we can represent the ARMA process as moving average model with infinitely many parameters.

3.7 Estimation of parameters in ARMA processes

In this section we present the most important results concerning the maximum likelihoood estimators, their confidence intervals and their asymptotic distributions. If data are not

Gaussian it is standard anyway to apply these confidence limits to the estimates even though they have been obtained through least squares.

The asymptotic distribution of the parameter vectors in an ARMA process are normally distributed. This means in particular that the paramaters themselves are normally distributed. The estimates and their standard deviations are produced by PEST. To obtain confidence intervals we only have to multiply these standard deviations with $\Phi_{1-\alpha/2}$, i.e. the $1-\alpha/2$ quantile in the $N(0,1)$ distribution, and add and subtract this new value from the estimate.

3.8 Order selection methods

In this section we will present some advanced order selection techniques. The problem of settling the right order of the ARMA process, or even AR or MA process, is far from trivial.

The oldest approach to order selection uses a method of estimating the variances of the white noise terms and plotting these values for different models. The idea is the following. If a model has too few parameters included, then the remaining random variation will be put into the error term, and the variance of the white noise will hence be bigger than the variance of the true error term. If, however, the model is overparametrized, then any further increase in parameters will not decrease the variance of the error term significantly, since we altready have enough parameters to explain the model. If we thus decide to plot the residual variance against the order of the model, then we will see a curve that first decreases for then later to level out. The point where it levels out could then indicate the right order of the model. Obviously this approach seems quite fragile, and partly subjective, and we will not recommend this approach in practice; at most it can serve as an additional check to the forthcomming more advanced methods.

In what follows we will describe the more advanced techniques known as Akaikes FTP, AIC, BIC and AICC indices.

3.8.1 The FTP index for AR processes

The ingenious idea of Akaike is basicly the following. Assume that the true order of the auto–regressive process is p. We let $\{X_t\}$ be the AR(p) process with parameters $a_1, ..., a_p$, i.e.

$$X_t + a_1 X_{t-1} + ... + a_p X_{t-p} = \epsilon_t,$$

and we let the white noise process have mean zero for convenience. Suppose $x_1, ..., x_n$ are data from the AR(p) process under consideration, and let $\hat{a}_1, ..., \hat{a}_p$ be the maximum likelihood estimates of the parameters obtained from these data.

Let $\{Y_t\}$ be an independent copy of the same AR(p) process. If we want to predict Y_{n+1} based on $Y_1, ..., Y_n$, then we can use the linear predictor

$$\hat{Y}_{n+1} = -a_1 Y_n - ... - a_p Y_{n+1-p}.$$

The mean square prediction error is then obviously

$$\mathbb{E}\left(\hat{Y}_{n+1} - Y_{n+1}\right)^2 = \mathbb{E}\epsilon_t^2 = \sigma_\epsilon^2.$$

The problem at this moment is that the coefficients $a_1, ..., a_p$ are unknown. If we, however, replace these coeffcients by their corresponding maximum likelihood estimates, then the mean square error is no longer σ_ϵ^2 but can be calculated to be

$$FTP(p) = \hat{\sigma}_\epsilon^2 \frac{n+p}{n-p},$$

where $\hat{\sigma}_\epsilon^2$ is the maximum likelihood estimate of the error term σ_ϵ^2.

The term FTP stands for "final prediction error", and when plotting $FTP(k)$ against k, the graph will in general show a well defined minimum. It is this value that we will use as an estimate for p.

But why choose the value at which the prediction error is at its minimum? At any time t, a value in an AR(p) process can be calculated from the values of the same process at time $t-1, ..., t-p$ plus an error. If we include too few parameters, say only $t-1, ..., t-q$, $q < p$, then too much variation will be put into the error term, and hence the maximum likelihood estimate of this error term will be too large in comparison with what it should have been. If on the other hand we use more parameters than necessary, i.e. prediction using the past times $t-1, ..., t-q$, where $q > p$, then the variation of variables that have no influence on the AR(p) process and its prediction is added, and this will consequently result in a larger prediction error. Thus the value at which FTP attains its minimum should be preferred.

3.8.2 AIC, BIC and AICC indices for ARMA processes

The Akiake AIC, BIC and AICC are indices all based on the same idea: to minimize a certain information criterion which puts costs on the number of variables. The more variables the more costly. The reason for putting costs on variables is the following. If data are originally from an AR(2) process, then of course any ARMA(p,q), with $p > 2, q \geq 0$ will fit the model equially well. And some of these models may even fit the data better. The reason for this is that we always have some random variation due to the error term, and some of this error may be removed by introducing additional variables. What the indices then do is to put a costs on these extra variables, so that we can distinguish

whether the improvement is significant or not. If the improvement is significant, the increase in the likelihood value will be bigger than the cost added.

Consider an ARMA(p,q) process given by

$$X_t + a_1 X_{t-1} + ... + a_p X_{t-p} = b_1 \epsilon_t + ... + b_q \epsilon_{t-q}.$$

Put $\boldsymbol{a} = (a_1, ..., a_p)'$ and $\boldsymbol{b} = (b_1, ..., b_q)'$, and their corresponding maximum likelihood estimates $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$. If L denotes the likelihood function then $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ are such that

$$\max L(\boldsymbol{a}, \boldsymbol{b}) = L(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}).$$

The AIC index is then defined as follows

$$AIC(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = -2 \log L(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) + 2(p + q + 1).$$

Thus we see that any increase in the likelihood will make the first term become smaller, but to make a significant better fit, the decrease has to be bigger than the increase from the second term.

The BIC index is given by

$$\begin{aligned} BIC &= (n - p - q) \log(n\hat{\sigma}_\epsilon^2/(n - p - q)) + n(1 + \log\sqrt{2\pi}) \\ &\quad + (p + q) \log(\frac{1}{p + q} \sum_{t=1}^{n} x_t^2 - n\hat{\sigma}_\epsilon^2). \end{aligned}$$

The AICC index is given by

$$AICC(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = -2 \log L(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) + 2(p + q + 1)n/(n - p - q - 2).$$

The differences between these indices is basically the following. The AIC index has a tendency to overestimate p, while the AICC index compensates for this by adding extra penalty for additional parameters. The index BIC is a consistent estimate for the true order, which means that as $n \to \infty$ the BIC index will converge to the true order, but may in practice well underestimate the true order. Consistency is not the case for neither AIC, AICC and FTP indices, but they are all more efficient estimates than the BIC index. This means in practice that we should put weight of preference to AICC, AIC, FTP and BIC in that order. All indices should however be taken into consideration when selecting a model, as well as the residual variance plots. A deviation from the expected in one of them may indicate something is wrong.

3.9 Order selection in practice

The order selection is a complex thing that should be learned in practice. Here, however, there is a rough guideline for things to be aware of.

First of all inspect carefully the graphs of the auto–corelation function and partial auto–correlation function. Do these suggest an AR model or MA model? One should at this point also be aware of the possibility of higher order AR or MA processes, where many parameters are set to zero. If e.g. the true model is $X_t + a_1 X_{t-1} + a_{12} X_{t-12} = \epsilon_t$, then the model only involves three parameters, but it is an AR(12) model. The partial auto–correlation function of this model is only non–zero at lags 1 and 12. If the estimated partial auto–correlation function shows such a behaviour it may be useful to try a higher–order model with many zero coefficients. Similar comments applied to MA processes, where the graph of the partial auto–correlation function is simply replaced by the graph of the auto–correlation function.

After a thorough inspection of these graphs, calculate AICC indices for the model ARMA(p,p), where $p = 1, 2, ...$ until minimum has been reached. The model attained at the minimum will form the basis for our further investigation. Consider the parameter estimates for this model, and more important, the parameter estimates divided by $1.96std.$, where *std* is the standard deviation of the parameter under consideration. According to (3.10), if the parameter value divided by $1.96std$ is between -1 and 1, then this is same as to say that a statistical test at level 95% for the hypothesis that the parameter value is zero is accepted, and similarly if the quotient is smaller than -1 or larger than 1 then we reject the hypothesis. These values can be used to eliminate parameters from the model, one by one, and every time a parameter has been eliminated we calculate the AICC index for the reduced model. If this is lower than the previous indices we shall consider this reduced model as a basis for further reduction. If not we proceed by trying to eliminate other parameters which quotients with $1.96std$ is between -1 and 1 as well.

Continuing this way we obtain a final model for our data. If this model is an ARMA model it should be compared to possible higher order AR, MA and ARMA models, if the graphs of the auto–correlation and partial auto–correlation functions of the residuals suggests so. If a higher order AR,MA or ARMA process is considered as well we compare the AICC indices of these models with that of the estimated lower order ARMA model.

When calculating the AICC index and other indices as well it is important to notice that we use the maximum likelihood approach. When performing the preliminary analysis the AICC index that appears there (in PEST) is based on Durbin–Levinson's algorithm, and is hence not a precise estimate for the true index value. The preliminary estimation is only used for creating initial values for the maximum likelihood procedure, and when comparing AICC indices only use those calculated by the maximum likelihood method.

At this point it is also appropriate to come with a warning for not misusing the

values created by the parameter estimate divided by $1.96std$. Suppose we are calculating the AICC indices of ARMA(p,p) models, and inspecting the estimated models. If the values created by dividing the parameters by $1.96std$ for some larger lags turns out to be insignificant (between -1 and 1) then this is not necessarily a sign that the true order of the model has been passed. This is only the AICC index that can decide this. For example, higher order models with many zero parameters will indeed have the feature that the parameter estimates divided by $1.96std$ will be insignificant for most parameters. This does, however, not imply that search should be stopped the first time an additional parameter turns out to be insignificant.

On the other hand, if we are searching only for a lower order ARMA model where there are no zeros among the parameters, then additional parameters with parameter estimates that divided by $1.96std$ is between -1 and 1 indicates that the true order has been passed. If the AICC index does not show the same behaviour (continue to drop), then this is an indication that there will be zeroes in the model, and that the there are parameters not yet included that will be significant.

3.10 Diagnostics

In this section we discuss various methods for checking the fit of the model to the data.

3.10.1 Stationarity

In case stationarity has not been obtained we cannot rely on the analysis we perform subsequently. Therefore a thorough check for stationarity has to be made.

We have already mentioned methods for how to transform data into stationary mode when certain periodicity or trends are available. In this section we mention some more detailed methods for checking stationarity.

First we have to consider the auto–correlation function of the data. If the series is non–stationary is has a tendency to converge slowly to zero, while in turn stationary processes usually have a auto–correlation function that dies out fast.

The visual inspection of the auto–correlation function of course requires some practical experience for how to decide what is 'slowly convergence' to zero. To obtain this experience it is recomendable for the inexperienced user to simulate some stationary ARMA processes of various orders, and consider the tail of the auto–correlation functions. These simulation should be compared to some real data sets which are obviously not stationary, like the airpass data.

A formal statistical test can also be performed using a regression method. Suppose that the data we want to test for stationarity is represented by the time series Y_t. Then regress ∇Y_t on $Y_{t-1} - \bar{Y}$, ∇Y_{t-1},...,∇Y_{t-p}. If the model under consideration is an AR(q)

model then put $p = q - 1$, or if it is an ARMA process, choose p large enough to make a good fit to the data.

Then the t–statistic for the parameter estimate of $Y_{t-1} - \bar{Y}$, defined in the usual way as the parameter estimate divided by its standard deviation, provides a test for nonstationarity. The distribution of this test statistic does, however, not have a t–distribution, but a distribution that has been simulated and listed in a table called τ_μ. Then compare the usual t–statistic with the critical values of the τ_μ–distribution.

3.10.2 Goodness of fit of the model

When a presumeably appropriate model has been selected it is important to check the goodness of fit of the model. To this end we inspect the residuals, and test for independence. If an ARMA model indeed fits the data, then the residuals have to be uncorrelated. Thus as a first inspection we plot the graphs of the auto–correlation function and partial auto–correlation function. Inserting the lines $y = 1.96/\sqrt{n}$ and $y = -1.96/\sqrt{n}$ which corresponds to end points in the 95 % confidence interval for the mean value of an i.i.d. sequence, 95 % of the points from the auto–correlation function and partial auto–correlation functions have to be between these two lines. If more values are significantly outside the confidence region, we have to reject our model.

3.10.3 Tests for white noise errors

The visual inspection of the plots of the auto–correlation function and partial auto–correlation function is only a part of the diagnostics, though an important one. We will in following list a sequence of tests for independence which will all be applied in the analysis.

The Portmanteau test

The test–statistic Q_W is used as follows. Choose a number (usually 20 or $\sqrt{n}$), and reject the hypothesis of independence at level α if

$$Q_W > \chi^2_{1-\alpha}(h - p - q).$$

This test is very generous to the non–stationary models in the sense that only estreme non–stationary models are rejected.

Turning points test

Independence is rejected if the test–statistic T satisfies

$$|T - \mu_T|/\sigma_T > \Phi_{1-\alpha/2},$$

where T in the program PEST is given by its asymptotic normal distribution

$$T \sim N(\mu_T, \sigma_T^2).$$

Difference sign test

Independence is rejected if the test–statistic S satisfies

$$|S - \mu_S|/\sigma_S > \Phi_{1-\alpha/2},$$

where S in the program PEST is given by its asymptotic normal distribution

$$S \sim N(\mu_S, \sigma_S^2).$$

The difference sign test will not reject data with a strong cyclic behaviour.

Rank test

Independence is rejected if the test–statistic P satisfies

$$|P - \mu_P|/\sigma_P > \Phi_{1-\alpha/2},$$

where P in the program PEST is given by its asymptotic normal distribution

$$P \sim N(\mu_P, \sigma_P^2).$$

3.11 The basic analysis

In this section we present a scheme that applies to the first important part of our analysis: the model selection, its estimates and some diagnostics.

I **Data transformations.**

- *Box–Cox transformation.* If the data show a increasing variablity through time, a Box–Cox transformation can be used make this variability more constant. In particular, if the standard deviation increases linearly over time the logarithmic transformation can be applied.
- *Seasonality.* If the data posseses a clear cyclic behaviour this can be removed either by the classical techniques or by applying a difference operator at the apropriate lag. The difference method should be preferred.
- *Trends.* Any polynomial trend of order k can be removed by k succesive applying the difference operator at lag one. Classical techniques are usually restricted to linear or quadratic trends. The difference method should again be preferred. Note that if a seasonal difference operator is applied it will remove a possible linear trend as well.

II **Preliminary estimation and order selection**

- *Mean adjustment.* Subtract the mean from the data, since this is a crucial assumption in our analysis developed.
- *Visual inspections.* Use the graphs of the auto–covariance and partial auto–correlation functions to inspect for whether the model is likely to be a pure MA, AR or a mixed model (ARMA).
- *Fit ARMA(p,p) models.* Fit ARMA(p,p) models in the following way. First apply preliminary estimation to obtain starting values for the maximum likelihood procedure. Then apply the maximum likelihood procedure to calculate the AICC index, and other indices as well. When the AICC index is at its minimum, choose the corresponding model to be the basis for further analysis. Call this model for the best ARMA(p,p) model.
- *Parameter elimination.* Use the values given by the parameter estimates divided by their respective standard errors times 1.96 as a guideline for parameter elimination. Select for example the numerically smallest such

value (i.e closest to zero) and replace the parameter by zero. Reestimate this new model, and calculate its AICC index. If this is smaller than the best ARMA(p,p), let this reduced model take its place as best ARMA(p,p) model and proceed as before. If not, try remove other parameters by the same procedure. Continue until no further parameters can be removed (using the AICC index).

- *Higher order models.* Include, if the graphs of the auto–correlation function or partial auto–correlation function suggests so, higher order AR, MA or ARMA models in the analysis. Compare with the estimated lower order ARMA process.

III **Diagnostics.**

- *Plot auto–correlations and partial auto–correlations on residuals.* All lags from one and up should be within the confidence lines. If one or more are outside, investigate if it significantly ouside in case of doubt. If so, the model has to be rejected.
- *Test for independence.* Apply various test for independence to the residuals.

Chapter 4

ARIMA and Multiplicative models

4.1 ARIMA processes

ARIMA processes is only a way of formulating (possibly non–stationary) models we have already been considering under ARMA models. A time series $\{X_t\}$ is called ARIMA(p,d,q) if $(1-B)^d X_t$ is a cuasal ARMA process. Here we of ourse assume that d is a non–negative integer.

This means that $\{X_t\}$ satisfies the difference equation

$$\alpha(B)(1-B)^d X_t = \beta(B)\epsilon_t.$$

If we call $\alpha_1(z) = \alpha(z)(1-z)^d$, then we see that α_1 has a root of multiplicity d at 1, and hence the ARIMA model is only stationary for $d = 0$.

ARIMA models hence includes time series that can be described as an ARMA model after a d–dimensional polynomial trend has been removed, as we have considered previously. But they are more than that. Even if there is no trend present, the model may be an ARIMA process if it shows obvious deviations from stationarity, for example a slowly deaying sample auto–correlation function. To find an appropriate level for d, apply the first order difference operator succesively until the sample auto–correlation function decreases rapidly, and hence makes it possible to fit a lower order ARMA process to the differenced data.

4.2 Roots and non–stationarity

Consider first a zero mean AR(k) process,

$$\alpha(B)X_t = \epsilon_t, \tag{4.1}$$

where $\alpha(z) = 1 + a_1 z + ... + a_k z^k$. Basically (4.1) is a difference equation, which general solution can written as the solution to the corresponding homogeneous equation plus a particular solution. We already know from (4.1) that a particular solution is given by $X_t = \alpha(B)^{-1}\epsilon_t$, wheras the homogeneous equation $\alpha(B)f(t) = 0$ has the general solution

$$f(t) = A_1\mu_1^t + ... + A_k\mu_k^t,$$

where $A_1, ..., A_k$ are constants and $\mu_1, ..., \mu_k$ are roots of the polynomial $g(z) = z^k + a_1 z^{k-1} + ... + a_k$. Note that if μ_i is a root of $g(z)$ then $\mu_i^k + a_1\mu_i^{k-1} + ... + a_k = 0$ and dividing by μ_i^k gives

$$1 + a_1\frac{1}{\mu_i} + ... + a_k\frac{1}{\mu_i^k} = \alpha(\frac{1}{\mu_i}) = 0.$$

Thus $1/\mu_i$ is a root in $\alpha(z)$.

To obtain asymptotic stationarity (as $t \to \infty$) we must require that $\lim_{t\to\infty} f(t) = 0$ but this is only the case if for all i, $|\mu_i| < 1$. This is again equivalent to say that all roots of $\alpha(z)$ have to outside the unit circle $\{z : |z| \leq 1\}$.

Concerning the general form of the auto–correlation function we use the Yule–Walker equations

$$\rho(m) + a_1\rho(m-1) + ... + a_k\rho(m-k) = 0, \quad m = k, k+1, ...$$

The general solution to this homogeneous equation is exactly like for $f(t)$, namely

$$\rho(r) = B_1\mu_1^r + ... + B_k\mu_k^r, \quad r \geq 0$$

for some constants $B_1, ..., B_k$, and symmetrical for negative r's. We see from this general solution that if $|\mu_i| < 1$ for all i, then $\rho(r)$ decay to zero as $z \to z^r$ as $r \to \infty$ for some z with $|z| < 1$. This means a very rapid decay in the tail, i.e for larger lags. If in turn not all roots of $\alpha(z)$ are outside the unit circle, then we are in the non–stationary case, and there exists a μ_i such that $|\mu_i| \geq 1$. Only in very pathological cases we can construct a non–stationary process where the covariance is increasing through time (corresponding to $|\mu_i| > 1$), and we will not consider such cases here. If μ_i is a point on the perifiri of the unit circle, i.e. $\mu_i = e^{i\theta}$ (where i in the exponential refers to the imaginary number and not the index i), then $|\mu_i| = 1$ and we see that $B_i\mu_i^r$ does not decay as $r \to \infty$.

There may, however, be other roots of $\alpha(z)$ outside the unit circle wich make contributions $B_j\mu_j^r$ that decays to zero, so all in all will $\rho(r)$ show a slowly decaying behaviour for smaller lags (say up to 100) for thereafter to level out and never converge to 0.

Thus non–stationarity can be seen on the estimated auto–correlation function as a slowly decaying behaviour, wheras in turn stationarity shows a fast decaying behaviour. This remark is very important in practice when we are checking for stationarity. If the

ACF estimate decays fast, this simply excludes the possibility of roots of $\alpha(z)$ inside the unit circle, and hence confirms stationarity. This procedure is therefore very safe to use, and by far better than any existing test for stationarity.

Same arguments applies to causal, invertible ARMA processes, since we can represent or approximate those with finite AR models.

4.3 Roots critically close to the unit circle

If the true underlying process we try to identify is an ARMA process where the polynomial $\alpha(z)$ has one or more roots close to but ouside the unit circle we may prefer to describe the model as a non–stationary ARIMA model rather than a stationary ARMA model. The reason for this is the following.

If $\alpha(z)$ did infact have a root in $\{z : |z| \leq 1\}$ then the ARMA model would not be causal, and causality is our overall assumption for stationarity: if a process is not causal, and hence cannot be expressed as a future independent general linear process (which is by definition stationary), then it fails to be among this very general class of stationary time series, and is then most likely to be non–stationary. Of course it might be that it is simply future dependent but stationary, but we would not like to consider that kind of processes, and hence we define processes that fail to be causal to be non–stationary.

If an ARMA process has a root very close to the unit circle, it may then in practice be impossible to disginguish the data from a non–stationary process. Consider for example the simplest case of an AR(1) model $X_t + aX_{t-1} = \epsilon_t$. If the true value of a is 0.99, then we will need a huge sample to be able to estimate a with such a precision that the confidence interval of the parameter does not include the value 1. Hence in this case, for normal data sizes, it is not possible to disginguish the stationary ARMA model with a non–stationary model. This leads to the conclusion that the estimation procedures may be very sensitive to models with a root close to the unit circle, since most estimation procedures break down when the model is non–stationary.

It is therefore desirable to transform the data in such a way that the roots are well ouside the unit circle. In theory this means that non of the roots have confidence regions that may hit the unit circle. In practice it means that the sample auto–correlation function decreases rapidly to zero.

ARIMA processes may be considered as a non–stationary ARMA processes (for $d > 0$) with a (multiple) root on the unit circle at the point (1,0) in the complex plane. ARIMA processes do not include the possibily that the polynoimal may have a root on the unit circle away from (1,0), i.e. in a point $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, where $\theta \in (-\pi, \pi]$ but different from 0. We will discuss this phenomenon in the following.

First notice that any (causal) ARMA process have an auto–correlation function that can be expressed in terms of the roots of the polynomial $\alpha(z)$, and therefore as well in

terms of the dual polynomial $g(z)$ that has roots which are the reciprocals of the roots of $\alpha(z)$. The roots of $g(z)$ are usually called $\mu_1, ..., \mu_p$. We don't need the details for how to write the auto–correlation function exact in terms of the roots, but the following intuitive argument shows that the statement is true. Write the ARMA process as

$$X_t = \frac{\beta(B)}{\alpha(B)}\epsilon_t.$$

The coeffcients g_u in the general linear representation (see ARMA processes, section 7) is then given by

$$\frac{\beta(z)}{\Pi_{i=1}^{p}(1-\mu_i z)} = \sum_{u=0}^{\infty} g_u z^u.$$

Then, in principle, we could find g_u by differentiating the left hand side u times and evaluate at 0. This would then obviously give a rational function depending on the parameters in $\beta(z)$, u and the roots of $g(z)$, μ_i. Then using the fact that the auto–correlation function is given by

$$\rho(h) = \sum_{u=-\infty}^{\infty} g_u g_{u-h} / \sum_{u=-\infty}^{\infty} g_u^2,$$

where $g_u = 0$ for $u < 0$, we have then proved that the auto–correlation function of an ARMA process is indeed a function of the roots of α.

After this intermezzo, we will now draw some striking conclusions using this fact. Let us suppose that we are confronted with a non–stationary process, and calculate the "auto–covariance" function $\mathbf{Cov}(X_1, X_h)$. Note that the auto–covariance function does of course not exist in this case in the sense that $\mathbf{Cov}(X_t, X_{t+h})$ depends on t, but we set $t = 1$ and calculate $\mathbf{Cov}(X_t, X_{t+h})$ for this value. Then if the non–stationarity is due to a root at complex number (1,0) (as for ARIMA processes) then the "auto–correlation" function we decay slowly as a rational function (a quotient between two polynomials, e.g. $1/x$, $(x^2+3)/(x^3+x^2+2x+1)$ for then to level out at larger lags.

In the case where the non–stationarity is due to root at $e^{i\theta}$ for $\theta \neq 0$, then the functions $\cos(\cdot)$ and $\sin(\cdot)$ will impose a oscillating behaviour of the "auto–correlation" function.

Thus we conclude, that non–stationarity can be seen on the auto–correlation function of the data as a slow decay in the ACF. In the case of a simple decrease (like a rational function) we may try to model the process by an ARIMA model, while if the auto–correlation function shows a oscillating behaviour the class of ARIMA models will be insufficient to model the data.

The same results applies to models which are stationary, but which have roots close to the unit circle; such models are best treated as non–stationary.

If a time series has a root at $e^{i\theta}$ then we cannot make it stationary by using the lag–1 difference operator ∇ succesively as for ARIMA processes. Such processes will necessarily

show a cyclic behaviour in the ACF (though it may be decreasing at the same time). If the root is $e^{i\theta}$ then we may look for an integer k such that $k\theta = 2\pi$ approximately. Then k is the length of the period of the cycles, and the operator $\nabla_k = 1 - B^k$ can then be applied succesively to the data to remove the non–stationarity. For more advanced techniques on detecting θ we refer to the forthcomming chapter on spectral analysis.

4.4 Multiplicative models

Also called seasonal ARIMA models or SARIMA models, multiplicative models allow for a more general trend and seasonality variation over time, which often is called "adaptive".

Fist let us define SARIMA models. A time series $\{X_t\}$ is called a SARIMA$(p,d,q) \times (P,D,Q)_s$ or, multiplicative (with the same parameters), with period s if $Y_t = (1-B)^d(1-B^s)^D X_t$ is a causal ARMA process on the form

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)\epsilon_t,$$

where ϵ_t is a zero mean white noise process with variance σ^2, $\phi(z)$, $\Phi(z)$, $\theta(z)$, and $\Theta(z)$ are polynomials of order p, P, q, Q respectively.

The parameter D is rarely more than 1. Thus is plays the role of a sort of indicator for if there is seasonality present ($D = 1$) or not ($D = 0$). Larger values for D than 1 may be justified by a ACF plot after applying the operator $(1 - B^s)$ already once.

One can show that if $\theta(B)\Theta(B^s)$ contain the factors $(1 - B)^d$, $(1 - B^s)^D$ then the seasonal component has a strictly stable "deterministic" form, and the trend has a stable "deterministic" polynomial form with constant coefficients. If on the other hand these factors are not contained in $\theta(B)\Theta(B^s)$ then the seasonal component has an adaptive form, where amplitudes and phases changes over time, and the trend has an adaptive polynomial form.

We can interpret the multiplicative model in the following way. Suppose we are considering a time series that is periodic with period 12. Then we file all values for january in a separate file, all values for february in a separate file etc. Each of these data files are then assumed to follow the same ARMA(P,Q) model, and can hence be written on the form

$$\Phi(B^{12})X_t = \Theta(B^{12})U_t,$$

where U_{j+12t} is the white noise process corresponding to the data from month j in each year. Furthermore we assume that the errors between months are correlated such that we obtain that the data from different data files are correlated as well. Suppose that the noise process satisfies the ARMA requirement

$$\phi(B)U_t = \theta(B)\epsilon_t,$$

where $\{\epsilon_t\}$ is a white noise process. Then we can write $U_t = \phi(B)^{-1}\theta(B)\epsilon_t$, and hence

$$\Phi(B^{12})X_t = \Theta(B^{12})\phi(B)^{-1}\theta(B)\epsilon_t,$$

or

$$\phi(B)\Phi(B^{12})X_t = \theta(B)\Theta(B^{12})\epsilon_t,$$

which is the original SARIMA formulation.

Thus in order to identify a multiplicative model it will be useful to use this interpretation. After having made stationary the original series by applying $(1-B)^d(1-B^s)^D$ to the data we proceed by finding the orders of P and Q by visual inspection of the ACF and PACF at lags that are multiples of s. We should choose P and Q in accordance with the actual form of the ACF and PACF of an ARMA(P,Q) model. Next we choose p and q by considering lags $1, 2, ..., s-1$ of the ACF and PACF. Finally we apply our standard techniques (AICC, parameter reduction techniques) to continue our analysis based on this model or more competing models.

Chapter 5

Frequency Analysis

5.1 Introduction

Formally we may consider a (deterministic) cycle as a sine or cosine wave. Consider e.g. the cosine wave

$$f_\nu(t) = \cos(2\pi\nu t).$$

Here ν plays the role of the frequency of the wave: the number of oscillation the wave will perform in a unit interval $[0, 1]$. If $\nu = 1$ the wave will perform one cycle, $\nu = 2$ two cycles etc.

The period T_0 is simply the reciprocal of the frequency, i.e.

$$T_0 = \frac{1}{\nu}.$$

If e.g. the frequency is $\nu = 10$ then there are 10 cycles per unit interval, and hence the period or length of each cycle is hence $\frac{1}{10}$.

Consider a time series on the form

$$x_t = \sum_{k=1}^{q} \left(a_k \cos(2\pi\nu_k t) + b_k \sin(2\pi\nu_k t)\right).$$

This time series model is a mixture of sine and cosine waves with different frequencies. Let us consider the simple example where $q = 2$, $a_1 = 1$, $a_2 = 0$, $b_1 = 0$, $b_2 = 0.75$, $\nu_1 = 0.0625$ and $\nu_2 = 0.2000$. Thus the model can be written as

$$x_t = \cos(2\pi(0.0625)t) + 0.75\sin(2\pi(0.2000)t).$$

This model is of course deterministic, but serves well as an illustration of our method. If in turn the coefficients a_k s and b_k s were random variable with zero mean mutually uncorrelated, and

$$\mathbb{E}(a_k^2) = \mathbb{E}(b_k^2) = \sigma_k^2,$$

then the series x_t is in fact stationary. The proof for this fact is a direct calculation of $\mathbb{E}(x_{t+m}x_t)$ and can be carried out by the reader (one should obtain $\sum_{i=1}^{q}\sigma_i^2\cos(2\pi\nu_k m)$).

The series x_t has a certain periodic behaviour composed of the two different frequencies $\nu_1 = 0.0625$ and $\nu_2 = 0.2000$. The small frequency makes the longer cycles, while the higher frequence makes "cycles on longer cycles".

From this simple example we would more or less be able to figure out the frequencies from this figure, but for more complex situations it may be desirable to have a procedure that can indicate these frequencies. Such a procedure is referred to as Fourier transform methods.

Consider the following transforms,

$$\begin{aligned} X_{\cos}(k) &= \frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} x_t\cos(2\pi\nu_k t) \\ X_{\sin}(k) &= \frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} x_t\sin(2\pi\nu_k t) \\ P_x(\nu_k) &= X_{\cos}^2(k) + X_{\sin}^2(k), \end{aligned}$$

where the frequencies are taken to be $\nu_k = k/T$. The first transform is called the cosine transform, the second the sine transform and the third is called the periodogram. Let us calculate the three transforms for the series x_t. We set $T = 16$. Then

$$\begin{aligned} X_{\cos}(k) &= \frac{1}{4}\sum_{t=0}^{15} x_t\cos(2\pi\nu_k t) \\ X_{\sin}(k) &= \frac{1}{4}\sum_{t=0}^{15} x_t\sin(2\pi\nu_k t) \\ P_x(\nu_k) &= X_{\cos}^2(k) + X_{\sin}^2(k), \end{aligned}$$

where $\nu_k = \frac{k}{16}$, $k = 0, 1, 2, ..., 8$. Note at this point, that it is usual only to specify these transforms for frequencies up to $T/2$. The reason for this will be apparent later when we see that the periodogram is symmetrical. Thus for a complete specification in $[-\pi, \pi]$ it is only needed to calculate values in $[0, \pi]$, which indeed correspond to values of ν_k from $k = 1, ..., [T/2]$, where $[\cdot]$ denotes the integer part. The values are the following:

k	ν_k	$X_{\cos}(k)$	$X_{\sin}(k)$	$P_x(\nu_k)$
0	0.0000	0.00	0.00	0.00
1	0.0625	2.01	0.06	4.04
2	0.1250	0.07	0.16	0.03
3	0.1875	0.75	1.12	1.82
4	0.2500	-0.29	-0.29	0.17
5	0.3125	-0.18	-0.12	0.04
6	0.3750	-0.15	-0.06	0.03
7	0.4375	-0.14	-0.03	0.02
8	0.5000	-0.14	0.00	0.02

The cosine transform has a peak at the frequence 0.1875, influenced from the real frequency of 0.2000, and the sine transform has a peak at 0.0625, which corresponds to the real frequency of the sine contribution. The periodogram comprises the cosine and sine transforms, and has peaks both at frequencies 0.0625 and 0.1875. Thus we obtained a sort of emperical fealing for how the transforms work in practice: the sine transform has peaks at frequencies that corresponds to sine contributions, the cosine transform at frequencies that corresponds to cosine terms, and the periodogram has peaks whenever a sine or cosine term is present with the corresponding frequency.

Moreover we notice the robustness of the transforms: the peak of the sine transform and periodogram at 0.1875 is influenced by the real frequency 0.2000. Thus in practice when observing a peak at a certain frequency we can never be sure that the frequency under consideration is the right one; in fact one can almost be sure it is not, but that the real frequency is close to the frequency at which we observe the peak.

5.2 Spectral densities

We now return to our stochastic analysis, and introduce some formal framework. Spectral analysis is a way of converting the auto–correlation function into a function that states something about the cycles in a stationary time series. Remember that cycles in a time series can also be seen as cycles in the ACF, and the spectral analysis tries to reveal these cycles and their frequencies.

Frist let $\{\rho(h)|h = 0, \pm 1, \pm 2, ...\}$ be any sequence of numbers. Then (Wold's theorem) $\rho(h)$ is the auto–correlation function of some (stationary) time series $\{X_t\}$ if and only if there exists a function F with the properties of a distribution function on the interval $(-\pi, \pi)$ (i.e. $F(-\pi) = 0, F(\pi) = 1$ and F is non–decreasing) such that

$$\rho(h) = \int_{-\pi}^{\pi} e^{i\omega h} dF(\omega).$$

If F is differentiable everywhere, then the normalized power spectral density function $f(\omega) = F'(\omega) = \frac{d}{d\omega}F(\omega)$ exists, and we have the following relation

$$\rho(h) = \int_{-\pi}^{\pi} e^{i\omega h} f(\omega) d\omega.$$

This expression can be inverted to the following Fourier series expansion

$$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \rho(h) e^{-i\omega h}, \quad -\pi \leq \omega \leq \pi.$$

For real-valued processes (which is the only kind of processes we consider in this course) we have that

$$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \rho(h) \cos(h\omega), \quad -\pi \leq \omega \leq \pi,$$

or

$$f(\omega) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{h=1}^{\infty} \rho(h) \cos(h\omega), \quad -\pi \leq \omega \leq \pi.$$

The function F may then be recovered by integration

$$F(\omega) = \int_{-\pi}^{\omega} f(\theta) d\theta = \frac{\omega + \pi}{2\pi} + \frac{1}{\pi} \sum_{h=1}^{\infty} \rho(h) \frac{\sin(h\omega)}{h}.$$

The spectral density $h(\omega)$ is based on the auto-covariance function $\gamma(h) = \sigma_X^2 \rho(h)$, and defined in a similar way as the normalized spectral density, namely

$$h(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) \cos(h\omega),$$

for $-\pi \leq \omega \leq \pi$.

Let $\{X_t\}$ be an ARMA(p,q) process with the usual representation

$$\phi(B) X_t = \theta(B) \epsilon_t,$$

where ϕ and θ have no common zeros and ϵ_t is a zero mean white noise with variance σ^2. Then the spectral density of X_t exists and is given by

$$h(\omega) = \frac{\sigma^2}{2\pi} \frac{|\theta(e^{-i\omega})|^2}{|\phi(e^{-i\omega})|^2}, \quad -\pi \leq \omega \leq \pi.$$

For obvious reasons the spectral density of an ARMA process is often called a rational spectral density.

5.3 Periodogram

The periodogram is a statistic that captures the features of the spectral density. In some books the periodogram is an attempt to estimate the spectral density, while in (most) others it is only proportional to an estimate of h. In practice this does not matter much since we are mostly interested in the frequencies where the density have peaks rather than the actual values of the density itself.

As in the introduction we define the periodogram, I_N, based on the observations $x_1, ..., x_N$ by

$$I_N(\omega) = A(\omega)^2 + B(\omega)^2,$$

where A and B are respectively the cosine and sine transforms,

$$\begin{aligned} A(\omega) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i \cos(\omega i) \\ B(\omega) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i \sin(\omega i). \end{aligned}$$

The periodogram can also be written as

$$I_N(\omega) = \frac{1}{N} \left| \sum_{t=1}^{N} x_t e^{-i\omega t} \right|^2.$$

To see this connection,

$$\begin{aligned} \left| \sum_{t=1}^{N} x_t e^{-i\omega t} \right|^2 &= \sum_{t=1}^{N} x_t e^{-i\omega t} \sum_{t=1}^{N} x_t e^{i\omega t} \\ &= \left(\sum_{t=1}^{N} x_t \cos(\omega t) - i \sum_{t=1}^{N} x_t \sin(\omega t) \right) \\ &\quad \times \left(\sum_{t=1}^{N} x_t \cos(\omega t) + i \sum_{t=1}^{N} x_t \sin(\omega t) \right) \\ &= \left(\sum_{t=1}^{N} x_t \cos(\omega t) \right)^2 + \left(\sum_{t=1}^{N} x_t \sin(\omega t) \right)^2 \\ &\quad + i \sum_{t=1}^{N} x_t \cos(\omega t) \sum_{t=1}^{N} x_t \sin(\omega t) \\ &\quad - i \sum_{t=1}^{N} x_t \cos(\omega t) \sum_{t=1}^{N} x_t \sin(\omega t) \\ &= N \left(A(\omega)^2 + B(\omega)^2 \right). \end{aligned}$$

Next we attempt to explain why the periodogram peaks when we hit or are near one of the "real" frequencies in the model. Let us consider the pure harmonic model

$$X_t = \sum_{i=1}^{K} (A_i \cos(\omega_i t) + B_i \sin(\omega_i t)) + \epsilon_t,$$

where we assume that the frequencies $\omega_i = 2\pi p_i/N$ for some integers $p_1, ..., p_K$, $0 \leq p_i \leq [N/2]$. This assumptions helps considerably in our calculations, and does not interact with the general argument which follows.

The harmonic model may be considered as a standard multiple linear regression model, where A_i, B_i are the unknown parameters. Thus we may apply a least squares method to estimate the unknown parameters. Put

$$Q = \sum_{t=1}^{N} \left(X_t - \sum_{i=1}^{K} (A_i \cos(\omega_i t) + B_i \sin(\omega_i t)) \right)^2.$$

Our task is then to minimize Q with respect to A_i and B_i. Differentiating Q w.r.t. A_j gives

$$\frac{d}{dA_j} Q = \sum_{t=1}^{N} 2 \left(X_t - \sum_{i=1}^{K} (A_i \cos(\omega_i t) + B_i \sin(\omega_i t)) \right) \cos(\omega_j t),$$

and solving for $\frac{d}{dA_j} Q = 0$ gives

$$\begin{aligned}
\sum_{t=1}^{N} X_t \cos(\omega_j t) &= \sum_{t=1}^{N} \sum_{i=1}^{K} (A_i \cos(\omega_i t) + B_i \sin(\omega_i t)) \cos(\omega_j t) \\
&= \sum_{i=1}^{K} A_i \sum_{t=1}^{N} \cos(\omega_i t) \cos(\omega_j t) + \sum_{i=1}^{K} B_i \sum_{t=1}^{N} \sin(\omega_i t) \cos(\omega_j t) \\
&= \sum_{i=1}^{K} A_i c_{ij} + \sum_{i=1}^{K} B_i d_{ij},
\end{aligned}$$

where $c_{ij} = \sum_{t=1}^{N} \cos(\omega_i t) \cos(\omega_j t)$ and $d_{ij} = \sum_{t=1}^{N} \sin(\omega_i t) \cos(\omega_j t)$. Similarly differentiating w.r.t. B_j and eveluating at zero gives

$$\sum_{t=1}^{N} X_t \sin(\omega_j t) = \sum_{i=1}^{K} A_i d_{ji} + \sum_{i=1}^{K} B_i s_{ij},$$

where $s_{ij} = \sum_{t=1}^{N} \sin(\omega_i t) \sin(\omega_j t)$. These equations are normal equations. Now using that the frequencies are given by

$$\omega_i = \frac{2\pi p_i}{N},$$

we remind the reader about the orthogonality relations

$$\sum_{t=1}^{N} \cos\left(\frac{2\pi pt}{N}\right)\cos\left(\frac{2\pi qt}{N}\right) = \begin{cases} 0 & 0 \le p \ne q \le [N/2] \\ N/2 & 0 < p = q < [N/2] \\ N & p = q = 0 \quad \text{or} \quad p = q = N/2,\ N \ \text{ even} . \end{cases}$$

$$\sum_{t=1}^{N} \sin\left(\frac{2\pi pt}{N}\right)\cos\left(\frac{2\pi qt}{N}\right) = 0$$

$$\sum_{t=1}^{N} \sin\left(\frac{2\pi pt}{N}\right)\sin\left(\frac{2\pi qt}{N}\right) = \begin{cases} 0 & 0 \le p \ne q \le [N/2] \\ N/2 & 0 < p = q < [N/2] \\ N & p = q = 0 \quad \text{or} \quad p = q = N/2,\ N \ \text{ even} . \end{cases}$$

Then we immediatly have that $c_{ij} = s_{ij} = 0$ for $i \ne j$, $d_{ij} = 0$ for all i, j and $c_{ii} = s_{ii} = N/2$ where i is such that $\omega_i \ne 0$ and $\omega_i \ne \pi$ (corresponding to $p \ne 0$ and $p \ne N/2$ when N is even). Thus the normal equations reduce to

$$A_i = \frac{2}{N}\sum_{t=1}^{N} X_t \cos(\omega_i t),\ B_i = \frac{2}{N}\sum_{t=1}^{N} X_t \sin(\omega_i t),$$

so as estimators for A_i and B_i we have that

$$\hat{A}_i = \frac{2}{N}\sum_{t=1}^{N} X_t \cos(\omega_i t),\ \hat{B}_i = \frac{2}{N}\sum_{t=1}^{N} X_t \sin(\omega_i t).$$

The mean of the estimators are resp. A_i and B_i, so they are unbiased. The variance of the estimators are given by

$$\mathbf{var}(\hat{A}_i) = \mathbf{var}(\hat{B}_i) = \frac{2\sigma_\epsilon^2}{N}.$$

We now relate the estimators $\hat{A}_i$ and $\hat{B}_i$ to the periodogram. If ω_p coincides with one of the real frequencies in the harmonic model ω_i (which we can assmue is different from 0 and $N/2$; if not we are in a trivial case) then we can write the cosine and sine transforms on the form

$$\begin{aligned} A(\omega_p) &= \frac{\sqrt{N}}{2}\hat{A}_i \\ B(\omega_p) &= \frac{\sqrt{N}}{2}\hat{B}_i \end{aligned}$$

Let

$$I_p = I_N(\omega_p), \quad \omega_p = 2\pi p/N,\ p = 0, 1, ..., [N/2].$$

Then

$$\mathbb{E}(I_p) = \frac{N}{4}\left(\mathbb{E}(\hat{A}_i^2) + \mathbb{E}(\hat{B}_i^2)\right).$$

But $\mathbb{E}(\hat{A}_i^2) = \mathbb{E}(\hat{A}_i)^2 + \mathbf{var}(\hat{A}_i) = A_i^2 + 2\sigma_\epsilon^2/N$ so, by a similar argument for $\mathbb{E}(\hat{B}_i^2)$, we get that

$$\mathbb{E}(I_p) = \frac{N}{4}\left(A_i^2 + B_i^2\right) + \sigma_\epsilon^2,$$

whenever $\omega_p = \omega_i$.

If on the other hand the frequency ω_p is far from any of the real frequencies ω_i we do the following trick. Insert the frequency ω_p in the model by putting $A_p = B_p = 0$. Then the same calculation as above results in

$$\mathbb{E}(I_p) = \sigma_\epsilon^2.$$

Thus we conclude that the expected value for the periodogram when $\omega_p = \omega_i$ is large (of order $O(N)$) while the expected value far from ω_i s is small (of size $O(1)$). Formally we are only making the destinction between if $\omega_p = \omega_i$ for some i or not. In practice, however, we also need to consider whether ω_p may be close to some ω_i or not. The reason is the following. If ω_p is indeed very close to some ω_i then it may be very difficult to distinguish ω_p from ω_i using a normal sized data set; thus the periodogram will still tend to show a peak for values ω_p close to ω_i. This feature is rather fortunate since it provides us with some robustness in estimating the frequencies at which there are peaks in the periodogram.

This little excursion into the world of harmonic processes explains in a heuristic manner the elementary features of the periodogram, that was already demonstrated in the introduction.

5.4 Tests for hidden periodicities

Testing goodness of fit of a time series model is often done by testing for white noise of the residuals. The Portmanteau test and the other non–parametric tests we have already considered, serve as basic tests for white noise. They are, however, unable to reveal more sophisticated trends such as cyclic or periodic behaviours. In this section we describe some tests that are able to detect periodicities; the null hypothesis is that the data are a Gaussian white noise and the alternative hypothesis is that the data are a Gaussian white noise added a periodic function. Thus in this context the "data" we are referring to will often be the residuals.

Let $\{X_t\}$ be our time series. Suppose that

$$X_t = \mu + A\cos(\omega t) + B\sin(\omega t) + \epsilon_t,$$

where A, B, ω are known constants, and ϵ_t is a Gaussian (normal) white noise with variance σ^2.

Then we want to test

$$H_0 : A = B = 0$$

against

$$H_1 : A, B \text{ are not both zero.}$$

We only carry out the test whenever $\omega = 2\pi k/n$ for some k such that $\omega \in (0, \pi)$. We reject H_0 in favour of H_1 at level α if

$$\frac{(n-3)I(\omega)}{\sum_{i=1}^{n} X_i^2 - I(0) - 2I(\omega)} > F_{1-\alpha}(2, n-3)$$

A more general tests is known as Fisher's test and goes as follows. We want to test the hypothesis $H_0 : X_1, ..., X_n$ is a Gaussian white noise against $H_1 : X_1, ..., X_n$ is a Gaussian white noise plus a deterministic periodic component. Here we do not need to specify the frequency or equivalently the exact period. Let $q = [(n-1)/2]$, and

$$\xi_q = \frac{\max_{1 \le i \le q} I(\omega_i)}{q^{-1} \sum_{i=1}^{q} I(\omega_i)},$$

where $\omega_i = 2\pi i/n$. If x denotes the observed value of ξ_q then we reject the null hypothesis at level α if

$$\mathbb{P}(\xi_q \ge x) = 1 - \sum_{j=0}^{q} (-1)^j \frac{q!}{j!(q-j)!} (1 - jx/q)_+^{q-1}$$

is less than α. Here the symbol $b = (a)_+$ means that $b = a$ whenever a is positive and zero otherwise.

In Fisher's test we conclude in case of rejection that there exist a periodic component of some unspecified frequency. From the test–statistic ξ_q we see that the method of the test is essentially to compare the maximum peak of the periodogram with its average level. The question then arises: can this unspecified frequency be another frequency than the one that corresponds to $\max I(\omega_i)$? It can be shown that the probability that the periodicity is due to some other frequency than the one that gives the maximum peak is less than α. Hence we assume, in case of rejection in Fisher's test, that the frequency under consideration is the one giving rise to the maximum peak.

5.5 Spectral analysis in practice

The periodogram $I_N(\omega)$ can be written as

$$I_N(\omega_m) = \sum_{k=-(N-1)}^{N-1} \hat{\gamma}(k) e^{-ik\omega_m},$$

where

$$\hat{\gamma}(k) = \frac{1}{N}\sum_{t=1}^{N-|k|}(x_t - \bar{x})(x_{t+|k|} - \bar{x})$$

is the usual estimate of the auto–covariance function, and $\omega_m = 2\pi m/N$ is a Fourier frequency with $m \neq 0$.

To see this connection we use that

$$\begin{aligned} I_N(\omega_m) &= \frac{1}{N}\left|\sum_{t=1}^{N} x_t e^{-i\omega_m t}\right|^2 \\ &= \frac{1}{N}\sum_{t=1}^{N} x_t e^{-i\omega_m t}\sum_{t=1}^{N} x_t e^{i\omega_m t} \\ &= \frac{1}{N}\sum_{t=1}^{N}\sum_{s=1}^{N} x_s x_t e^{-i\omega_m t} e^{i\omega_m s} \\ &= \frac{1}{N}\sum_{t=1}^{N}\sum_{s=1}^{N}(x_s - \bar{x})(x_t - \bar{x})e^{-i\omega_m t} e^{i\omega_m s}, \end{aligned} \tag{5.1}$$

where the last relation follows from the orthogonality relations

$$\sum_{t=1}^{N} e^{i\omega_m t} = \sum_{t=1}^{N} e^{-i\omega_m t} = 0.$$

At this point it is most important that the frequency under consideration is a Fourier frequency. In fact most of the spectral analysis in practice relies on this assumption. But then from (5.1) the result follows by change of variable $k = s - t$.

This result then suggests that a good estimate for the spectral density would be

$$\hat{h}(\omega_m) = \frac{1}{2\pi} I_N(\omega_m).$$

This is, however, not the case! The variance of the periodogram $\mathbf{var}(I_N(\omega_m))$ does not converge to zero as $N \to \infty$, which proves that I_N can never be a consistent estimate of h (i.e. $I_N(\omega_m)$ does not converge to $h(\omega_m)$ as $N \to \infty$). Moreover,

$$\mathbf{Cov}(I_N(\omega_{m_1}), I_N(\omega_{m_2}))$$

decreases as N increases. This fact results in a erratic and wild behaviour of the periodogram, and since spectral densities are functions that are smooth in their behaviour it is useless to use the periodogram as an estimator for the spectral density.

The periodogram itself is however far from useless; it should only be used in the right way. If the data e.g. are normally distributed the periodogram $I_N(\omega_m)$ is the

maximum likelihood estimate of the spectral density h! This proves the optimality of the periodogram as an estimator of h. But how can this fact be true when at the same time the periodogram is wild and erratic, and fails to capture the similarity of a theoretical spectral density? The point is that $I_N(\omega_m)$ is an optimal estimate for each single m, point estimates, but as an estimate for the whole function h it fails. This feature is also known from many other branches of statistics, and in particular from density estimation theory.

The optimality and sufficiency of the periodogram suggest that in order to obtain a good estimator for h we should consider a function of the periodogram. To this end we may introduce the so called lag–window λ_N, which is a real even function by which we multiply the auto–covariance estimate. Thus we will estimate h in the following way

$$\hat{h}(\omega_m) = \frac{1}{2\pi} \sum_{s=-(N-1)}^{N-1} \lambda_N(s)\hat{\gamma}(s)\cos(s\omega_m).$$

We attach the suffix N to the lag–window to highlight the dependence on N. For example we may chose the lag–window to be the function

$$\lambda_N(s) = \begin{cases} 1 & |s| \leq M \\ 0 & |s| > M \end{cases}$$

where $M < N - 1$.

This function truncates the auto–covariance function for larger lags, and the idea behind this is that the auto–covariance function is very badly estimated in the tail, since the estimate is based on very few observations. Thus for an appropriate choice of M one might expect that λ_N has a smoothing effect on the periodogram. This is indeed the case, as we can see in the following argument.

We have the following representation of I_N,

$$I_N(\omega_m) = \sum_{k=-(N-1)}^{N-1} \hat{\gamma}(k)e^{-ik\omega_m}$$

which can be inverted by means of Fourier theory to

$$\hat{\gamma}(k) = \int_{-\pi}^{\pi} I_N(\omega)e^{ik\omega}d\omega.$$

Then inserting this expression into the expression for $\hat{h}(\omega_m)$ we get that, using cos, λ_N and $\hat{\gamma}$ are even functions,

$$\hat{h}(\omega_m) = \frac{1}{2\pi} \sum_{s=-(N-1)}^{N-1} \lambda_N(s)\hat{\gamma}(s)\cos(s\omega_m)$$

$$
\begin{aligned}
&= \frac{1}{2\pi} \sum_{s=-(N-1)}^{N-1} \lambda_N(s)\hat{\gamma}(s)e^{-i\omega_m s} \\
&= \frac{1}{2\pi} \sum_{s=-(N-1)}^{N-1} \lambda_N(s) \int_{-\pi}^{\pi} I_N(\theta)e^{ik\theta}d\theta e^{-is\omega_m} \\
&= \int_{-\pi}^{\pi} I_N(\theta)W(\omega_m - \theta)d\theta,
\end{aligned}
$$

where

$$
W(\theta) = \frac{1}{2\pi} \sum_{s=-(N-1)}^{N-1} \lambda_N(s)e^{-is\theta}.
$$

Thus we see that the spectral estimate is an integral of the periodogram weighted with the smoothing function W. The smoothing function W is called the spectral window.

The conclusion is that multiplying the auto–covariance function with a lag–window has the effect of smoothing the periodogram with a spectral window.

For example for the truncated lag–window

$$
\lambda_N(s) = \begin{cases} 1 & |s| \leq M \\ 0 & |s| > M \end{cases}
$$

we have the corresponding spectral window

$$
W(\theta) = \frac{1}{2\pi} \sum_{s=-M}^{M} \cos(s\theta) = \frac{1}{2\pi} \frac{\sin((M+1/2)\theta)}{\sin(\theta/2)},
$$

which is also know as the Dirichlet kernel.

A more commonly used lag–window is the Bartlett window, which is defined by

$$
\lambda_N(s) = \begin{cases} 1 - |s|/M & |s| \leq M \\ 0 & |s| > M \end{cases}
$$

The spectral window corresponding to this lag–window is

$$
W(\theta) = \frac{1}{2\pi M} \left(\frac{\sin(M\theta/2)}{\sin(\theta/2)} \right)^2.
$$

Various other windows have been proposed, but the important thing to notice for the applied user is that it does not matter too much which window is chosen, but it is indeed very important that some window is applied.

When a window is chosen, one has to choose M as well. There is no fixed rule for how to do this, but it should be noted that in order to obtain consitent estimates of the

spectral density one should choose M such that $M \to \infty$ as $N \to \infty$. Thus a fixed proportion like 20 % or 30% of N is certainly a possibility.

Another fact that should be noticed is the following. As M increases, so does the variance of the estimates, but their bias decreases. On the other hand, if M decreases so does the variance but the bias increases. This fact should also be taken into account when choosing M. What are we interested in? Low bias or low variance on the estimators?

Finally we will describe the aliasing effect of spectral analysis. This has to do with the problem of discrete sampling; and any time series we consider in these notes can be considered as a discrete sample of some phenomenon.

Let the phenomenon be f_t, say, in continuous time, and let us only sample at time $t = 1, 2, 3, ..., N$. Suppose f_t is a periodic function with 1.5 periods per unit time. Thus our sample $f_1, f_2, ..., f_N$ will also show a periodic behaviour but not with period 1.5. If we e.g. sample at time $t = 2, 4, 6, ...$, then again we may find a periodic behaviour of the sample, but with a period that is different from the original function or from the first sample.

This problem that a discrete sample may not capture the true frequencies of the underlying model is called the aliasing effect. The aliasing effect has to be studied carefully when a continuous record of data is available, and we are left with the job to sample from this continous record (could e.g. temperatures during the day, the exchange rate at any point of the day), but in most cases the data are already given and nothing can be changed.

One should though be aware of the fact that the frequencies measured/estimated from the data may not reflect that frequencies in the underlying phenomenon. The frequencies are, however, the true frequencies for the data sample, and since we are only left with a data sample in practice, all we can do is to fit a model using all the information we can possibly extract from the data. The fitted model may then refelct some features of the underlying true model, or it may not.

Chapter 6

Multivariate Time Series and Transfer models

6.1 Introduction

For many analyses or investigations in practice, the data are drawn from various sources of information. For example an electricity company may compare data of total current production of electricity with various other factors, such as industrial production, consumption of certain goods or average daily outdore temperatures. In order to draw conclusions, or to make decent predictions, it may be necessary, or at least desirable for the general understanding, to process the data analysis as a multivariate case, i.e. the data are not considered on a isolated basis, but always in relation to each other.

If the data of the electricity plant were X_{t1}, X_{t2}, X_{t3} and X_{t4} for respectively total electricity production , industrial production, consumption and temperature, we could then set up our time series analysis by considering the time series data given by the vector

$$\boldsymbol{X}_t = (X_{t1}, X_{t2}, X_{t3}, X_{t4})'.$$

An aim of our analysis could be to analyze the connection between the four components. In particular one might wish to test for independence of certain variables of each other, or to express one of the marginal processes as a linear combination of another.

We denote the vector in boldface to distinguish from the coordinates of the vector. In the following all vectors and matrices will be denoted in boldface. All vectors are by default column vectors; thus we may write the vectors as transposed of row vectors, which is denoted by the prime $\boldsymbol{X}'$.

Let $\boldsymbol{X} = (X_1, ..., X_m)'$ be a random vector. Then the mean of the vector is defined as the vector of the means,

$$\boldsymbol{\mu} = \mathbb{E}\boldsymbol{X} = (\mathbb{E}X_1, ..., \mathbb{E}X_m)'.$$

Moreover, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two random vectors, then the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by

$$\mathbf{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left((\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y})'\right).$$

6.2 Stationarity

As in the univariate case, the key assumption for multivariate time series is stationarity. Let $\boldsymbol{X}_t$ be a m–dimensional multivariate time series with $\mathbb{E}X_{ti}^2 < \infty$ for all t and i. This assumption ensures the existence of the covariance function. Let $\boldsymbol{\mu}_t = \mathbb{E}\boldsymbol{X}_t$, and let $\boldsymbol{\Gamma}(t+h,t) = \mathbf{Cov}(\boldsymbol{X}_{t+h}, \boldsymbol{X}_t)$. Then we say that $\boldsymbol{X}_t$ is (second order) stationary if and only if $\boldsymbol{\mu}_t$ and $\boldsymbol{\Gamma}(t+h,t)$ do not depend on t. In that case we define $\boldsymbol{\mu} = \boldsymbol{\mu}_t$ and $\boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}(t+h,t)$.

Assume that $\boldsymbol{X}_t$ is a m–dimensional stationary time series. If we let $\boldsymbol{\Gamma}(h) = \{\gamma_{ij}(h)\}$, then we see that $\gamma_{ij}(h) = \mathbf{Cov}(X_{(t+h)i}, X_{tj})$. In particular, if $i = j$ then $\gamma_{ij}(h)$ is the auto–covariance function of the i'th coordinate process X_{ti}. For $i \neq j$ we say that $\gamma_{ij}(h)$ is the cross covariance function between X_{ti} and X_{tj}. Note that in general $\gamma_{ij}(h) \neq \gamma_{ji}(h)$. Due to stationarity we have, however, that

$$\mathbf{Cov}(X_{(t+h)i}, X_{tj}) = \mathbf{Cov}(X_{ti}, X_{(t-h)j})$$

from which we conclude that

$$\boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}(-h)'.$$

The correlations are defined in the usual way,

$$\rho_{ii}(h) = \gamma_{ii}(h)/\gamma_{ii}(0),$$

and the cross–correlations are defined by

$$\rho_{ij}(h) = \frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}.$$

In particular, $\rho_{ij}(0)$ is the correlation between X_{ti} and X_{tj}.

As building blocs for ARMA models etc. we also need to introduce multivariate white noise processes. A process $\boldsymbol{Z}_t$ is a m–dimensional white noise with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ if $\boldsymbol{\Gamma}(h) = \boldsymbol{\Sigma}$ if $h = 0$, and 0 otherwise. In that case we write $\{\boldsymbol{Z}_t\} \sim WN(\mathbf{0}, \boldsymbol{\Sigma})$.

If for all $i, j = 1, ..., m$, $\sum_{h=-\infty}^{\infty} |\gamma_{ij}(h)| < \infty$ then $\boldsymbol{X}_t$ has a spectral density $\boldsymbol{h}$ that can be written as

$$\boldsymbol{h}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \boldsymbol{\Gamma}(h),$$

and which can be inverted into

$$\boldsymbol{\Gamma}(h) = \int_{-\pi}^{\pi} e^{i\lambda h}\boldsymbol{h}(\lambda)d\lambda.$$

Notice that since the cross-covariances are not in general symmetrical, the cross–spectral densities $f_{ij}(\lambda)$ are in general complex valued.

6.3 Estimation

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be observations from a m–dimensional time series. As in the univariate case we will estimate the mean vector by

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

The i'th coordinate of this estimator is simply the usual univariate estimate of the mean of the i'th process.

As in the univariate case we have consistency of this estimator under the mild conditions,

$$\mathbb{E}\left((\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})'(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})\right) \to 0$$

as $n \to \infty$ if $\gamma_{ii}(h) \to 0$ as $h \to \infty$. Moreover,

$$n\mathbb{E}\left((\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})'(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})\right) \to \sum_{i=1}^{m}\sum_{h=-\infty}^{\infty} \gamma_{ii}(h)$$

if $\sum_{h=-\infty}^{\infty} |\gamma_{ii}(h)| < \infty$ for all $i = 1, ..., m$.

We also have asymptotic normality in case $\boldsymbol{X}_t$ is expressable as an infinite moving average process; the result is, however, of limited inmportance since it does not provide any concrete information about how to obtain confidence limits.

To that end we simply approximate the confidence limits by applying the confidence limit for each univariate process.

To estimate $\boldsymbol{\Gamma}(h)$ we use the natural estimates

$$\hat{\boldsymbol{\Gamma}}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(\boldsymbol{X}_{t+h} - \bar{\boldsymbol{X}}_n)(\boldsymbol{X}_t - \bar{\boldsymbol{X}}_n)'$$

if $0 \leq h \leq n-1$, and

$$\hat{\boldsymbol{\Gamma}}(h) = \frac{1}{n}\sum_{t=-(h-1)}^{n}(\boldsymbol{X}_{t+h} - \bar{\boldsymbol{X}}_n)(\boldsymbol{X}_t - \bar{\boldsymbol{X}}_n)'$$

if $-(n-1) \leq h \leq 0$. The reason for the separate definitions for h being positive and negative is that we do not have the property $\boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}(-h)$ as in the univariate case.

Once estimated $\boldsymbol{\Gamma}$, and hence all $\gamma_{ij}(h)$, we estimate the cross-correlation function by

$$\hat{\rho}_{ij}(h) = \hat{\gamma}_{ij}(h)/(\hat{\gamma}_{ii}(h)\hat{\gamma}_{jj}(h))^{1/2},$$

which for $i = j$ simply is the auto–correlation function of the i'th process.

The consistency of $\hat{\boldsymbol{\Gamma}}$ can be proved under the condition that $\boldsymbol{X}_t$ is a general linear process, and the white noise process consist of independent random vectors.

The following result is of particular importance. Consider two univariate time series X_{t1} and X_{t2}, and suppose that

$$X_{t1} = \sum_{j=-\infty}^{\infty} \alpha_j Z_{t-j,1}, \quad X_{t2} = \sum_{j=-\infty}^{\infty} \beta_j Z_{t-j,2},$$

where $\{Z_{t,i}\}$, $i = 1, 2$ are two sequences of i.i.d random noises, that are also independent of each other. This implies that X_{t1} and X_{t2} are independent of each other. Then if $\sum_j |\alpha_j| < \infty$ and $\sum_j |\beta_j| < \infty$ then for $h \geq 0$,

$$\hat{\rho}_{12}(h) \sim AN\left(0, \frac{1}{n}\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)\right),$$

where AN stands for asymptotically normal.

This results has as main consequence the following: If we wish to test for independence of X_{t1} and X_{t2} then we have to know $\rho_{11}(h)$ and $\rho_{22}(h)$ for all h. These could in principle be substituted by estimated values, but there is a risk that we would obtain a poor estimate due to few correlation values, or poorly estimated correlation values.

Instead we use a trick that is called prewhitening. Prewhitening basically consists of transforming X_{t1} and X_{t2} into white noise processes (from there the name prewhitening, since the white noise process has a spectrum that corresponds to white light) by applying a linear filter, i.e. to express Z_{ti} as a (infinite) linear combination of the X_{ti}'s,

$$Z_{ti} = \sum_{j=0}^{\infty} \pi_0^{(j)} X_{t-j,i}.$$

This approach is not feasible in practice since the true model is hardly ever known. Instead we fit ARMA(p,q) models to X_{t1} and X_{t2}, and use the residuals from the two fits as our prewhitened processes.

In these two residual processes the $\rho_{11}(h) = \rho_{22}(h) = 0$ for $h \neq 0$, and $\rho_{11}(0) = \rho_{22}(0) = 1$. Thus the result above states that asymptotically $\hat{\rho}_{12}(h)$ for the prewhitened processes

is normally distributed with mean zero and variance $1/n$. Thus the 95% confidence limites are easily constructed as $\pm 1.96/\sqrt{n}$. Then plotting $\hat{\rho}_{12}(h)$ of the residual processes against h provides a test for independence of the two processes X_{t1} and X_{t2}: if 95% of the points are within the significance limits we accept the hypothesis of independence, and otherwise not.

To obtain the effect of prewhitening, namely that the sum $\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)$ reduces to 1, it is in fact enough to prewhiten only one of the processes. If we prewhiten X_{t1} only, then $\rho_{11}(h) = 0$ for all $h \neq 0$, which is enough to ensure that the sum equals 1.

As in the univariate case, significant values of $\hat{\rho}_{12}(h)$ for some lag h suggests such values included in the model. But in the multivariate case we should remember to prewhiten one of the series before we calculate $\hat{\rho}_{12}(h)$; if not we cannot conclude anything from the plot of $\hat{\rho}_{12}(h)$.

6.4 Multivariate ARMA processes

The definition of multivariate ARMA processes is similar to that of univariate processes. Let $\{\boldsymbol{X}_t\}$ be a m–dimensional stationary time series. Then we say that $\{\boldsymbol{X}_t\}$ is an ARMA(p,q) process if

$$\boldsymbol{\Phi}(B)\boldsymbol{X}_t = \boldsymbol{\Theta}(B)\boldsymbol{Z}_t,$$

where $\boldsymbol{\Phi}(z) = \boldsymbol{I} + \boldsymbol{\Phi}_1 z + ... + \boldsymbol{\Phi}_p z^p$, $\boldsymbol{\Theta}(z) = \boldsymbol{I} + \boldsymbol{\Theta}_1 z + ... + \boldsymbol{\Theta}_q z^q$, $\boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_p$, $\boldsymbol{\Theta}_1, ..., \boldsymbol{\Theta}_q$ are $m \times m$ matrices, $\boldsymbol{I}$ is the identity matrix, and $\{\boldsymbol{Z}_t\}$ is a white noise process $WN(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Causality and invertibility is defined in the same manner: the process $\boldsymbol{X}_t$ is causal if it can be expressed as a future independent moving average process (possibly of infinite order), and it is called invertible if the process has an (possibly infinite) autoregressive representation.

It turns out that a criterion for cuasality is

$$\det\boldsymbol{\Phi}(z) \neq 0,$$

for all $z : |z| \leq 1$ (z complex). Similarly a invertibility criterion is

$$\det\boldsymbol{\Theta}(z) \neq 0,$$

for all $z : |z| \leq 1$.

If $\boldsymbol{X}_t$ is causal, then by definition

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{C}_j \boldsymbol{Z}_{t-j},$$

and since $\boldsymbol{\Phi}(B)\boldsymbol{X}_t = \boldsymbol{\Theta}(B)\boldsymbol{Z}_t$ then we can find the coefficient matrices $\boldsymbol{C}_j$ by

$$\sum_{j=0}^{\infty} \boldsymbol{C}_j z^j = \boldsymbol{\Phi}^{-1}(z)\boldsymbol{\Theta}(z)$$

for $|z| \leq 1$. Similarly for invertible models we have per definition that

$$\sum_{j=0}^{\infty} \boldsymbol{D}_j \boldsymbol{X}_{t-j} = \boldsymbol{Z}_t$$

and we can find the coefficient matrices by

$$\sum_{j=0}^{\infty} \boldsymbol{D}_j z^j = \boldsymbol{\Theta}(z)^{-1}\boldsymbol{\Phi}(z)$$

for $|z| \leq 1$.

In practice one can apply the following algorithm to calculate the coefficient matrices:

$$\begin{aligned} \boldsymbol{C}_0 &= \boldsymbol{I} = \boldsymbol{D}_0 \\ \boldsymbol{C}_j &= \sum_{i=1}^{j} \boldsymbol{\Phi}_i \boldsymbol{C}_{i-j} + \boldsymbol{\Theta}_j \\ \boldsymbol{D}_j &= -\sum_{i=1}^{j} \boldsymbol{\Theta}_i \boldsymbol{D}_{j-i} - \boldsymbol{\Phi}_j \end{aligned}$$

where we define $\boldsymbol{\Theta}_j = 0$ for $j > q$ and $\boldsymbol{\Phi}_i = 0$ for $i > p$.

6.5 Estimation of ARMA models

The estimation methods are basically the same as for univariate models, though a serious complication is present for multivariate models. The basic estimation procedure is the maximum likelihood method, which superimposes that the data are sampled from a multivariate normal distribution. In the univariate case this procedure works without complications. In the multivariate case, however, there is a problem in the estimation of mixed ARMA processes (i.e. both containing autoregressive and moving average parameters).

The problem is that of identifiability, and which results from the likelihood surface does not uniquely define a maximum. Thus we may choose several matrices as parameters for our ARMA model that all optimizes the likelihood function.

This problem of non–uniqueness does not happen if we are considering models that are either pure auto–regressive or moving average. For that reason many computer packages

does not deal with multivariate ARMA processes at all. The ITSM package deals with multivariate AR processes only, and SAS deals with multivariate processes in a state–space setting.

If we absolutly do have to use a mixed ARMA model, one way of trying to "direct" the algorithm to the "true" maximum likelihood estimate is to use the maximum likelihood estimates of the univariate processes as starting values for our multivariate iteration. This does not guarantee, however, that the limit of our iteration will be the "true" maximum likelihood estimator of the data. And even worse: there are no ways to measure if we found the "true" model or not. For exactly this reason it is advisable to proceed with care in practical applications, and mainly to use pure auto–regressive or pure moving average models.

A simplification of the estimation procedure for multivariate AR processes is based on an extention of the Durbin–Levinson algorithm. In the univariate case this was our preliminary estimation procedure for AR processes. For multivariate processes we will use this procedure instead of the maximum likelihood estimation procedure in order to save computer time. As we know from the univariate case the estimates using the Durbin–Levinson algorithm are fairly precise, and the likelihood surface in the multivariate case is so complicated that the iterative maximum likelihood estimates may well end in a local maximum instead.

6.6 Coherence and Phase–spectra

Consider a multivariate (stationary) time series $\boldsymbol{X}_t$ of dimension two. All arguments in the following carry through immediately to higher dimensions.

Then we can write

$$\boldsymbol{X}_t = \begin{pmatrix} X_{t1} \\ X_{t2} \end{pmatrix}.$$

The covariance function $\boldsymbol{\Gamma}(h)$ of $\boldsymbol{X}_t$ can then be decomposed as

$$\boldsymbol{\Gamma}(h) = \begin{pmatrix} \gamma_{11}(h) & \gamma_{12}(h) \\ \gamma_{21}(h) & \gamma_{22}(h) \end{pmatrix}. \tag{6.1}$$

The functions $\gamma_{ii}(h)$, $i = 1, 2$ are the auto–covariance function of the marginal processes X_{ti} respectively, and $\gamma_{ij}(h)$, $i \neq j$, $i, j = 1, 2$ are the cross-covariance functions, which we know in general do not coincide with each other.

The spectral density $\boldsymbol{h}$ is given by

$$\boldsymbol{h}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \boldsymbol{\Gamma}(h),$$

which then by (6.1) can be written as

$$\boldsymbol{h}(\lambda) = \begin{pmatrix} h_{11}(\lambda) & h_{12}(\lambda) \\ h_{21}(\lambda) & h_{22}(\lambda) \end{pmatrix}. \tag{6.2}$$

In (6.2) the diagonal entries are simply the spectral densities of the respective marginal processes X_{t1} and X_{t2}. The functions $h_{12}(\lambda)$ and $h_{21}(\lambda)$ are the so–called cross–spectral densities, or simply cross spectra, of X_{t1} and X_{t2}. From (6.1) it is clear that

$$h_{12}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \gamma_{12}(h).$$

Since $\boldsymbol{\Gamma}(h)$ is not in general symmetrical, the cross–spectra are in general complex valued. Note also that $\gamma_{12}(h)$ and $\gamma_{21}(h)$ contain equivalent information since they are related by the property $\boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}(-h)^t$, so that we may write $\gamma_{12}(h) = \gamma_{21}(-h)^t$. Therefore also the cross–spectra h_{12} and h_{21} contain equivalent information, and it is sufficient to consider one of them.

Let us consider h_{12}. Since $h_{12}(\lambda)$ is in general complex valued we can write

$$h_{12}(\lambda) = c_{12}(\lambda) - iq_{12}(\lambda),$$

where $c_{12}(\lambda)$ is the real part of $h_{12}(\lambda)$ and $q_{12}(\lambda)$ is the negative of the imaginary part. In this context $c_{12}(\lambda)$ is called the co–spectrum, and $q_{12}(\lambda)$ the quadrature spectrum.

We may also express $h_{12}(\lambda)$ in polar coordinates,

$$h_{12}(\lambda) = \alpha_{12}(\lambda) e^{i\phi_{12}(\lambda)}.$$

Here $\alpha_{12}(\lambda)$ is obviously the length of $h_{12}(\lambda)$, and is consequently given by

$$\alpha_{12}(\lambda) = \sqrt{c_{12}^2(\lambda) + q_{12}^2(\lambda)}.$$

This length is referred to as the amplitude spectrum. The function $\phi_{12}(\lambda)$ is obviously the angle between $h_{12}(\lambda)$ and the first axis. This angle can also be calculated using that the proportion $-q_{12}(\lambda)/c_{12}(\lambda)$ is $\tan(\phi_{12}(\lambda))$, so that

$$\phi_{12}(\lambda) = \tan^{-1}(-q_{12}(\lambda)/c_{12}(\lambda)).$$

This angle is referred to as the phase–spectrum. The complex coherence $w_{12}(\lambda)$ is defined by

$$w_{12}(\lambda) = \frac{h_{12}(\lambda)}{\sqrt{h_{11}(\lambda) h_{22}(\lambda)}},$$

and the coherency is defined as the absolute value of this complex number, $|w_{12}(\lambda)|$. The coherency is a sort of measure of linearity between X_{t1} and X_{t2}. In fact it can be shown that the coherency is a correlation coefficient between random coefficients of components of X_{t1} and X_{t2} that have frequency λ. A linear relationship is indicated if the coherency is 1. In fact, if X_{t1} and X_{t2} have the linear relation

$$X_{t2} = \sum_{j=-\infty}^{\infty} \phi_j X_{t-j,1},$$

then the coherency between the two processes is 1, where $\sum_j |\phi_j| < \infty$. In this context we say that X_{t2} is given as a time–invariant filter of X_{t1}. It also follows that any linear filter applied to both processes do not change the coherency between them. Finally we remark that if both processes are uncorrelated then obviously the coherency is 0.

For regression models, where we in addition to the linear relationship also have added a noise, the coherence is not 1. This is due to the noise factor. Thus the linear gression $X_{t,1} = X_{t,2} + \epsilon_t$ have coherence 1 if and only if $\epsilon_t = 0$.

An interesting class of models to consider are the so–called delayed regression models, where we have

$$X_{t,1} = aX_{t-d,2} + \epsilon_t.$$

Here we have the cross–covariance function given by

$$\gamma_{12}(h) = \mathbf{Cov}(X_{t+h,1}, X_{t,2}) = a\gamma_{22}(h-d),$$

and cross–spectral density

$$\begin{aligned} h_{12}(\lambda) &= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \gamma_{12}(h) \\ &= \frac{a}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \gamma_{22}(h-d) \\ &= a e^{-i\lambda d} h_{22}(\lambda). \end{aligned}$$

The auto–covariance function of X_{t1} is given by

$$\begin{aligned} \gamma_{11}(h) &= \mathbf{Cov}(X_{t,1}, X_{t+h,1}) \\ &= \mathbf{Cov}(aX_{t-d,2} + \epsilon_t, aX_{t+h-d,2} + \epsilon_{t+h}) \\ &= a^2\gamma_{22}(h) + \mathbf{Cov}(\epsilon_t, \epsilon_{t+h}) \end{aligned}$$

which immediatly results in

$$h_{11}(\lambda) = a^2 h_{22}(\lambda) + h_\epsilon(\lambda),$$

where $h_\epsilon(\lambda)$ is the (constant) spectral density of the white noise process.

Using the expression for $h_{12}(\lambda)$ only we get that $c_{12}(\lambda) = a\cos(\lambda d)h_{22}(\lambda)$, $q_{12}(\lambda) = a\sin(\lambda d)h_{22}(\lambda)$, $\alpha_{12}(\lambda) = ah_{22}(\lambda)$ and $\phi_{12}(\lambda) = -\lambda d$. To calculate the coherency we use the relation between h_{11} and h_{22} to get

$$w_{12}(\lambda) = \frac{ah_{22}(\lambda)e^{-i\lambda d}}{\sqrt{h_{22}(\lambda)(a^2h_{22}(\lambda) + h_\epsilon(\lambda))}},$$

and hence

$$|w_{12}|(\lambda) = \frac{1}{\sqrt{1 + h_\epsilon(\lambda)/(a^2h_{22}(\lambda))}}.$$

The important thing to notice in this example is how we can detect the delay parameter d, namely by using the phase–spectrum $\phi_{12}(h)$: if there is a time delay the phase–spectrum is a linear function of the frequencies, and the slope of this linear function is the order of magnitude for the time delay.

6.7 Estimation of the cross–spectrum

Let us for simplicity of notation consider a two–dimensional time series $\boldsymbol{X}_t = (X_{t1}, X_{t2})'$. All arguments presented immediately carries over to higher order models.

Given observations $X_1, ..., X_n$. Then the periodogram $\boldsymbol{I}_n$ is defined by

$$\boldsymbol{I}_n(\omega_j) = \sum_{k=-(n-1)}^{n-1} \hat{\boldsymbol{\Gamma}}(k)e^{-ik\omega_j},$$

where $\omega_j = 2\pi j/n$ is any non–zero (Fourier) frequency. As in the univariate case, the periodogram is simply defined as the spectral density by replacing $\boldsymbol{\Gamma}$ by its estimator.

The periodogram is, again, a poor estimator for the spectral density, and it may be desirerable to smooth the periodogram with a lag–window to obtain a consistent estimator (i.e. an estimator that converges to the true spectral density function as the number of data increases to infinity). There is no reason, however, that the lag–window for two marginal processes should be chosen to be the same, or that the lag–window for the cross–spectrum should coincide with any of the lag–windows from the marginal processes. One should choose the lag–window such that it is useful for the purpose we consider. For example, if we consider the problem of estimating the coherency or phase–spectrum, then we should only be concerned with choosing a lag–window that is useful for smoothing the cross–spectrum, and if we are only interested in the marginal spectra we may need to estimate possibly twice with two different lag–windows, each of one useful for one of the marginals.

Suppose that we have chosen a lag–window $\lambda_n(s)$ useful for our purposes. Then we estimate the spectral density by

$$\hat{\boldsymbol{h}}(\omega_j) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \lambda_n(k) \hat{\boldsymbol{\Gamma}}(k) e^{-ik\omega_j}.$$

The theory for multivariate spectral analysis now basically follows the same track as for univariate spectral analysis, and the practical considerations are the same. We refer to the univariate theory for more practical details.

Notice that since $\boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}(-h)'$ we have that $h_{12}(\lambda)$ and $h_{21}(\lambda)$ are complex conjugates. This implies that the real part of $\boldsymbol{h}$ can be written as

$$c_{12}(\omega) = \mathrm{Re}\boldsymbol{h}(\omega_j) = \frac{1}{2}(h_{12}(\omega_j) + h_{21}(\omega_j)),$$

and the negative of the imaginary part as

$$q_{12}(\omega) = -\mathrm{Im}\boldsymbol{h} = \frac{1}{2} i(h_{12}(\omega_j) - h_{21}(\omega_j)).$$

We will use these relations to estimate c_{12} and q_{12} simply be replacing the cross–sepctral densities by their estimators.

The amplitude spectrum can now be estimated by

$$\hat{\alpha}_{12}(\omega_j) = \sqrt{\hat{c}_{12}(\omega_j)^2 + \hat{q}_{12}(\omega_j)^2}.$$

The phase–sepctrum is estimated by

$$\tan^{-1}(\hat{c}_{12}(\omega_j)/\hat{q}_{12}(\omega_j)).$$

Last, the complex coherency is straighforward to estimate replacing the spectra by its estimated spectra in the expression for the complex coherency. For estimation of the absolute coherency, or only coherency, we use

$$\frac{\hat{\alpha}_{12}(\omega_j)}{\sqrt{\hat{h}_{11}(\omega_j)\hat{h}_{22}(\omega_j)}}.$$

Approximate 95 % confidence intervals are given by the following formulae. Let

$$a_n = \sum_{|k| \leq m} \lambda_n(k)^2,$$

where m is the truncation point of the lag–window, i.e. the largest integer for which the lag–window is different from zero.

The phase–spectrum has confidence limits

$$\hat{\phi}_{12}(\omega_p) \pm 1.96 a_n \hat{\alpha}_{12}(\omega_p) \sqrt{(|\hat{w}_{12}(\omega_p)|^{-2} - 1)/2}.$$

The confidence interval for the coherence is

$$|\hat{w}_{12}| \pm 1.96 a_n (1 - |\hat{w}_{12}(\omega_p)|^2)/\sqrt{2}.$$

If the lag–window is the truncated window with

$$\lambda_n(k) = \frac{1}{2m+1}$$

for $|k| \leq m$, and 0 otherwise, we can test for independence in the following way. The null hypothesis is $|w_{12}(\omega)| = 0$ and the alternative $|w_{12}(\omega)| > 0$. The null hypothesis is accepted at level α if

$$\frac{2m|\hat{w}_{12}(\omega)|^2}{1 - |\hat{w}_{12}(\omega)|^2} \leq F_{1-\alpha}(2, 4m).$$

6.8 Transfer function modelling

The considerations in the previous section naturally leads to considerations about the possibility of expressing one time series as a linear combination of another, using the cross spectral properties. In this section we consider a special class of such models which are called transfer function models.

6.8.1 Basic formulation and analysis

Let $\{X_{t1}\}$ and $\{X_{t2}\}$ be zero mean stationary processes, and suppose they are related by

$$X_{t2} = \sum_{j=0}^{\infty} t_j X_{t-j,1} + N_t, \tag{6.3}$$

where $\{N_t\}$ is the noise processes, assumed to be a zero mean stationary time series, and uncorrelated with X_{t1}. Then the process X_{t1} serves as input process and X_{t2} serves as output process.

The objective of our analysis is to determine the coefficients t_j, $j = 0, 1, ...$ and the noise processes N_t for given data x_{t1} and x_{t2}. To determine the noise process is easy if we know the coefficients, since it is simply a residual process between X_{t1} and a linear filter applied to X_{t2}. Thus the whole problem basically boils down to calculating/estimating the coefficients t_j.

Multiply (6.3) by $X_{t-k,1}$ and obtain

$$X_{t2}X_{t-k,1} = \sum_{j=0}^{\infty} t_j X_{t-j,1}X_{t-k,1} + N_t X_{t-k,1}. \tag{6.4}$$

Since X_{t1} and N_t are uncorrelated we get

$$\begin{aligned} \mathbb{E}(N_t X_{t-k,1}) &= \mathbb{E}(N_t)\mathbb{E}(X_{t-k,1}) \\ &= 0 \end{aligned}$$

by the zero means assumption. Thus taking expectations in (6.4) yields

$$\gamma_{21}(k) = \sum_{j=0}^{\infty} t_j \gamma_{11}(k-j). \tag{6.5}$$

To estimate the t_j's from this equation would require precise estimation of the $\gamma_{11}(k)$ for all k. This is an unnecessary complication, and instead we use the concept of pre–whitening the input series X_{t1}. For if X_{t1} were indeed a white noise process the right hand side of (6.5) boils down to $t_k\sigma_1^2$, where σ_1^2 is the variance of X_{t1}.

Suppose X_{t1} is given by an ARMA(p,q) model,

$$\phi(B)X_{t1} = \theta(B)Z_t$$

where Z_t is a white noise process with zero mean and variance σ_Z^2. Define the filter $\pi(B)$ by

$$\pi(B) = \theta(B)^{-1}\phi(B) = \phi(B)\theta(B)^{-1}.$$

Applying this filter to X_{t1} will prewhiten the process. Indeed from $\phi(B)X_{t1} = \theta(B)Z_t$ we have that $\phi(B)\theta(B)^{-1}X_{t1} = Z_t$ or $\pi(B)X_{t1} = Z_t$. The method looks alright in theory, but how do we in practice apply this (non–linear) filter to our input process? Quite simple: estimating the ARMA model in the usual manner will produce residuals which are the filtered process we are looking for. Let $Y_t = \pi(B)X_{t2}$ be the same filter applied to X_{t2}. How do we do that in practice? Use the estimated ARMA coefficients of the X_{t1}, $\hat{\phi}_1, ..., \hat{\phi}_p$, $\hat{\theta}_1, ..., \hat{\theta}_q$ to determine a process Z_t' such that

$$\hat{\phi}(B)X_{t2} = \hat{\theta}(B)Z_t'.$$

Then $Y_t = Z_t'$. Another way of formulating this concept: use $\hat{\phi}_1, ..., \hat{\phi}_p$, $\hat{\theta}_1, ..., \hat{\theta}_q$ as initial values for the maximum likelihood estimation procedure, and apply the maximum likelihood estimation with 0 iterations, and calculate the residuals from this model.

Applying the filter $\pi(B)$ to N_t, using similar methods as above, results in a new stationary process N'_t. Thus after applying the filter $\pi(B)$ to (6.3) we get the equation

$$Y_t = \sum_{j=0}^{\infty} t_j Z_{t-j} + N'_t.$$

From this equation we can easily calculate the coeffcients by multiplying both sides of the equation with Y_{t-k} and taking expectations. We get that

$$t_k = \gamma_{YZ}(k)/\sigma_Z^2.$$

Now

$$\rho_{YZ}(k) = \frac{\gamma_{YZ}(k)}{\sqrt{\sigma_Z^2 \sigma_Y^2}}$$

so

$$t_k = \rho_{YZ}(k)\sigma_Y/\sigma_Z.$$

To use the cross–correlation functions has the following practical advantage. If we plot $\hat{\rho}_{YZ}(k)$ against the confidence limits $\pm 1.96/\sqrt{n}$ then we can get a first impression of which t_j are zero and which are not. Note that $t_j = 0$ if and only if $\rho_{YZ}(j) = 0$. The non–zero values are then estimated by

$$\hat{t}_j = \hat{\rho}_{YZ}(j)\hat{\sigma}_Y/\hat{\sigma}_Z.$$

For non–signinficant values of $\hat{\rho}_{YZ}(j)$ we put $t_j = 0$, and force the parameters at such lags to remain zero. The smallest value such that $\rho_{YZ}(h)$ is non–zero, b, is called the delay parameter of the filter $\{t_j\}$, and the largest such value, the order of the transfer model, is denoted by m. Now we have classified significant and non–significant values of t_j, and we have proposed some preliminary estimators for these values. By this ends our use of the prewhitened series, and we return to our original stationary series X_{t1} and X_{t2}.

We proceed by analysing the noise process N_t, which we extract by

$$\hat{N}_t = X_{t2} - \sum_{j=b}^{m} \hat{t}_j X_{t-j,1}. \tag{6.6}$$

Then fit an ARMA model to the noise process $\hat{N}_t$ to obtain

$$\phi^{(N)}(B)\hat{N}_t = \theta^{(N)}(B)W_t, \tag{6.7}$$

where $\{W_t\}$ is a white noise process with zero mean and variance σ_W^2. This fit will be used when we specify our final model in the following section.

6.8.2 Parameter reduction and extention of the model

Let $T(B) = \sum_{j=b}^{m} t_j B^j$. Then from (6.6) and (6.7) it follows that

$$\phi^{(N)}(B)X_{t2} = T(B)\phi^{(N)}(B)X_{t1} + \theta^{(N)}(B)W_t. \tag{6.8}$$

Here $T(B) = B^b(t_b + t_{b+1}B + ... + t_m B^{m-b})$ is a simple polynomial. If m is large (say arround 7–8 and up) we may succesfully apply an approximation to T of an rational function with fewer parameters. In the final formulation of our model we will extend T to be on the form

$$T(B) = B^b \frac{w_0 + w_1 B + ... + w_s B^s}{1 - v_1 B - ... - v_t B^t} = B^b \frac{w(B)}{v(B)}.$$

This function of course has as special case the polynomial form of T by choosing $t = 0$ and $w_i = t_{i+b}$. If the sequence of $\hat{t}_j$ s decay approximately like

$$\hat{t}_j = w_0 v_1^{j-b}$$

for $j \geq b$ then we may choose

$$T(B) = \frac{w_0}{1 - v_1 B} B^b.$$

Appropriate values for w_0 and v_1 are found using the first two preliminary estimates of $\hat{t}_j$.

Using the extended function T and dividing through by $\phi^{(N)}(B)$ in (6.8) we get that

$$X_{t2} = B^b \frac{w_0 + w_1 B + ... + w_s B^s}{1 - v_1 B - ... - v_t B^t} X_{t1} + \frac{\theta^{(N)}(B)}{\phi^{(N)}(B)} W_t. \tag{6.9}$$

This model will serve as our model specification for our tansfer model. Thus in order to chose the rational function

$$B^b \frac{w_0 + w_1 B + ... + w_s B^s}{1 - v_1 B - ... - v_t B^t}$$

we use the preliminary estimates of $\hat{t}_j$, with the possibility of choosing a polynomial ($t = 0$), which coincides with the original and exact model formulation. Choosing a rational function ($t > 0$) will only serve as an approximation to a higher order polynomial. Note that if we use a polynomial approximation, m is no longer well defined (it is significantly larger than 0 for all lags) and we put $m = \max(s + b, t)$.

More advanced, if the series $\hat{t}_j$ satisfies the difference equation

$$\hat{t}_j - v_1 \hat{t}_{j-1} - ... - v_t \hat{t}_{j-t} = 0 \tag{6.10}$$

then T is well approximated by

$$\hat{T}(B) = \frac{w_0 B^b}{1 - v_1 B - ... - v_t B^t}.$$

Whether the t_j s satisfies a difference equation like (6.10) may be difficult to check in practice, but a possibility could be to estimate an AR process with small white noise variance and use the estimates of the parameters as the coefficients $v_1, ..., v_t$. The order t should as well be estimated by the AICC criterion applied to the series $\hat{t}_j$.

Now we have specified parameters for all entries in the model (6.9), and we will use these parameters as preliminary estimates for initiating a least squares fit to the model. The least squares procedure will optimize the parameters $\hat{t}_j$. Then calculate a new residual noise process N_t using these optimized coefficients, and fit an ARMA model to these residuals. If the order of the ARMA model is the same as before we have finished; if not, fit a new ARMA model with different orders, and repeat the least squares fitting with the new polynomial $\phi^{(N)}$ and $\theta^{(N)}$ included in the model (6.9). Continue in this way until the order of ARMA model fitted to the residual noise process we obtain from the optimized parameters is the same as the order of the previous fitted ARMA model.

Diagnostics for transfer function models simply consists of checking the appropriateness of all ARMA models estimated using the usual diagnostic techniques, and to check for uncorrelatedness between W_t and the input series. This is done by plotting the cross-correlations between W_t and the prewhitened input process, i.e. between W_t and the residuals of the input process from its ARMA estimation. If uncorrelated the values of the cross–correlation should be insignificant.

6.9 Intervention analysis

Intervention analysis is an important application of transfer models to univariate data that contain an obvious change of level at a certain point. Such data will never appear stationary becuase of the change of level, and even though we can model the data in the following way.

Suppose the data are X_t and that there is an obvious change of level at time $t = c$. By applying the usual techniques to the data to obtain stationarity, the most we can hope for is to obtain two stationary series that are stationary before time c and after time c. The full series will never appear stationary.

Construct a series Y_t that is 0 if $t < c$ and 1 if $t \geq c$. Consider the transfer function model

$$X_t = aY_t + N_t.$$

We may think of N_t as the original series X_t without a change of level at time c. The term aY_t compensates for the change of level, since it adds the constant a for all values of

$t \geq c$. When applying stationarity transformations to X_t, in order to obtain two stationary sequences (one up to time c and after time c), we apply the same operations to the right hand side. If the difference operator ∇ is involved we see that Y_t reduces to a series which is zero everywhere but for one or two values (two values if an operator ∇_k, $k > 1$ has been applied as well to remove e.g. a seasonal component). When differencing the level–changing series, we obtain two series with equal fluctuations, but in the changing point there will be a heavy fluctuation.

Thus we may describe our transfer model in terms of stationary series as

$$X_{t1} = aY_{t1} + N_{t1},$$

where X_{t1} is the series X_t after stationarity transformations, and Y_{t1} and N_{t1} the series corresponding to Y_t and N_t after the same stationarity transformations. The interpretation of this equation is the following: N_{t1} corresponds to a stationary sequence without a heavy fluctuation at $t = c$. But by adding a only at this value $t = c$ we can simulate this "boost" and apply standard transfer modelling to the data. The only other alternative would be to throw away the data before time c, and this is obviously not recomendable since the data before time c contain equally valid information on the process, as the points after c.

One may ask whether the analysis is valid, since the input process is obviously not a random process, and can hardly, at least from an outstanding point of view, be regarded as stationary. These points of critics are indeed valid, and the only argument in our favor is that the procedure works fine in practice. We can actually estimate an MA(1) or AR(1) with $\phi_1 = 0$ or $\theta_1 = 0$ and a small variance to the differenced data of Y_t (at least if ∇ has been involved, and there are only on point different from 0).

If there are more than one level changing we may a apply a transfer model that contains as many non–zero t_j's as there are level changes in the data. Then every non–zero t_j corresponds to a "heavy" fluctuation occuring at some point.

Bibliography

The reader who whishes to aquintain himself with further details and proofs for the material presented in these notes may consult Priestly (1981), Brockwell & Davis (1991) and Anderson (1971). Priestley (1981) provides an excellent treatment of the spectral analysis problems involved in time series analysis, and provide many useful practical hints. Brockwell and Davis (1991) is a more theoretical exposition, but very readable and much recommended for the reader who wants to put the theory of time series into a more mathematical frame, such as Hilbert spaces. Anderson (1971) is a classic book which for that reason still serves as a reference. It is also worth mentioning Brillinger (1981) and Hannan (1970). For further references, in particular to articles and specialized expositions, see references in Priestley (1981) or Brockwell & Davis (1991). Finally, for a discussion on prediction methods, see Makridakis et al. (1984).

References

Anderson, T.W. (1971) *The statistical analysis of time series.* John Wiley, New York.

Brockwell, P.J. & Davis, R.A. (1991) *Time series: theory and methods.* Springer Verlag, New York, Berlin.

Hannan, E.J. (1970) *Multiple time series.* John Wiley, New York.

Makridakis et al. (1984) *The forecasting accuracy of major time series methods.* John Wiley, New York.

Priestley, M.B. (1981) *Spectral analysis and time series.* Academic Press, New York.

SERIE DOCUMENTOS DE TRABAJO

The following working papers from recent year are still available upon request from:

Rocío Contreras,
Centro de Documentación, Centro de Estudios Económicos, El Colegio de México A.C., Camino al Ajusco # 20 C.P. 01000 México, D.F.

90/I Ize, Alain. "Trade liberalization, stabilization, and growth: some notes on the mexican experience."

90/II Sandoval Musi, Alfredo. "Construction of new monetary aggregates: the case of Mexico."

90/III Fernández, Oscar. "Algunas notas sobre los modelos de Kalecki del ciclo económico."

90/IV Sobarzo, Horacio E. "A consolidated social accounting matrix for input-output analysis."

90/V Urzúa, Carlos M. "El déficit del sector público y la política fiscal en México, 1980 - 1989."

90/VI Romero, José. "Desarrollos recientes en la teoría económica de la unión aduanera."

90/VII García Rocha, Adalberto. "Note on mexican economic development and income distribution."

90/VIII García Rocha, Adalberto. "Distributive effects of financial policies in Mexico."

90/IX Mercado, Alfonso and Taeko Taniura "The mexican automotive export growth: favorable factors, obstacles and policy requirements."

91/I Urzúa, Carlos M. "Resuelve: a Gauss program to solve applied equilibrium and disequilibrium models."

91/II Sobarzo, Horacio E. "A general equilibrium analysis of the gains from trade for the mexican economy of a North American free trade agreement."

91/III Young, Leslie and José Romero. "A dynamic dual model of the North American free trade agreement."

91/IV Yúnez-Naude, Antonio. "Hacia un tratado de libre comercio norteamericano; efectos en los sectores agropecuarios y alimenticios de México."

91/V Esquivel, Hernández Gerardo. "Comercio intraindustrial México-Estados Unidos."

91/VI Márquez, Colín Graciela. "Concentración y estrategias de crecimiento industrial."

92/I Twomey, J. Michael. "Macroeconomic effects of trade liberalization in Canada and Mexico."

92/II Twomey, J. Michael. "Multinational corporations in North America: Free trade intersections."

92/III Izaguirre Navarro, Felipe A. "Un estudio empírico sobre solvencia del sector público: El caso de México."

92/IV Gollás, Manuel y Oscar Fernández. "El subempleo sectorial en México."

92/V Calderón Madrid, Angel. "The dynamics of real exchange rate and financial assets of privately financed current account deficits"

92/VI Esquivel Hernández, Gerardo. "Política comercial bajo competencia imperfecta: Ejercicio de simulación para la industria cervecera mexicana."

93/I Fernández, Jorge. "Debt and incentives in a dynamic context."

93/II Fernández, Jorge. "Voluntary debt reduction under asymmetric information."

93/III Castañeda, Alejandro. "Capital accumulation games."

93/IV Castañeda, Alejandro. "Market structure and innovation a survey of patent races."

93/V Sempere, Jaime. "Limits to the third theorem of welfare economics."

93/VI Sempere, Jaime. "Potential gains from market integration with individual non-convexities."

93/VII Castañeda, Alejandro. "Dynamic price competition in inflationary environments with fixed costs of adjustment."

93/VIII Sempere, Jaime. "On the limits to income redistribution with poll subsidies and commodity taxation."

93/IX Sempere, Jaime. "Potential gains from integration of incomplete markets."

93/X Urzúa, Carlos M. "Tax reform and macroeconomic policy in Mexico."

93/XI Calderón, Angel. "A stock-flow dynamic analysis of the response of current account deficits and GDP to fiscal shocks."

93/XII Calderón, Angel. "Ahorro privado y riqueza financiera neta de los particulares y de las empresas en México."

93/XIII Calderón, Angel. "Política fiscal en México."

93/XIV Calderón, Angel. "Long-run effects of fiscal policy on the real levels of exchange rate and GDP."

93/XV Castañeda, Alejandro. "On the invariance of market innovation to the number of firms. The role of the timing of innovation."

93/XVI Romero, José y Antonio Yúnez. "Cambios en la política de subsidios: sus efectos sobre el sector agropecuario."

94/I Székely, Miguel. "Cambios en la pobreza y la desigualdad en México durante el proceso de ajuste y estabilización".

94/II Calderón, Angel. "Fiscal policy, private savings and current account deficits in Mexico".

94/III Sobarzo, Horacio. "Interactions between trade and tax reform in Mexico: Some general equilibrium results".

94/IV Urzúa, Carlos. "An appraisal of recent tax reforms in Mexico". (Corrected and enlarged version of DT. Núm. X-1993)

94/V Sáez, Raúl E. and Carlos Urzúa. "Privatization and fiscal reform in Eastern Europe: Some lessons from Latin America".

94/VI Feliz, Raúl. "Terms of trade and labour supply: A revision of the Laursen-Metzler effect".

94/VII Feliz, Raúl and John H. Welch. "Cointegration and tests of a classical model of inflation in Argentina, Bolivia, Brazil, Mexico, and Peru".

94/VIII Sempere, Jaime. "Condiciones para obtener ganancias potenciales de liberalización de comercio".

94/IX Sempere, Jaime y Horacio Sobarzo. "La descentralización fiscal en México: Algunas propuestas".

94/X Sempere, Jaime. "Are potential gains from economic integration possible with migration?".

94/XI Gollás, Manuel. "México 1994. Una economía sin inflación, sin igualdad y sin crecimiento".

95/I Schettino, Macario. "Crecimiento económico y distribución del ingreso".

95/II Schettino, Macario. "A function for the Lorenz curve".

95/III Székely P., Miguel. "Economic Liberalization, Poverty and Income Distribution in Mexico".