Time Series Analysis

Loglinear
Publishing

Michael Sampson

Terms of Use

This document is distributed "AS IS" and with no warranties of any kind, whether express or implied.

Until November 1, 2001 you are hereby given permission to print one (1) and only one hardcopy version **free of charge** from the electronic version of this document (i.e., the pdf file) provided that:

1. The printed version is for your personal use only.
2. You make no further copies from the hardcopy version. In particular no photocopies, electronic copies or any other form of reproduction.
3. You agree not to ever sell the hardcopy version to anyone else.
4. You agree that if you ever give the hardcopy version to anyone else that this page, in particular the **Copyright Notice** and the **Terms of Use** are included and the person to whom the copy is given accepts these **Terms of Use**.

Until November 1, 2001 you are hereby given permission to make (and if you wish sell) an **unlimited number of copies on paper only** from the electronic version (i.e., the pdf file) of this document or from a printed copy of the electronic version of this document provided that:

1. You agree to pay a royalty of either **$3.00 Canadian** or **$2.00 US** per copy to the author within 60 days of making the copies or to destroy any copies after 60 days for which you have not paid the royalty of **$3.00 Canadian** or **$2.00 US** per copy. Payment can be made either by cheque or money order and should be sent to the author at:

   Professor Michael Sampson
   Department of Economics
   Concordia University
   1455 de Maisonneuve Blvd W.
   Montreal, Quebec
   Canada, H3G 1M8

2. If you intend to make **five or more copies**, or if you can reasonably expect that five or more copies of the text will be made then you agree to notify the author **before** making any copies by **Email** at: **sampson@loglinear.com** or by fax at **514-848-4536**.
3. You agree to include on each paper copy of this document and at the same page number as this page on the electronic version of the document: 1) the above **Copyright Notice,** 2) the **URL**: **http://www.loglinear.com** and the **Email** address **sampson@loglinear.com**. You may then if you wish remove this **Terms of Use** from the paper copies you make.

Contents

Chapter 1

Trends, Cycles and Seasonality

1.1 Introduction

In time series analysis we study the probabilistic laws which determine how an economic time series W_t (such as real GDP, the level of employment or the $S\&P$ 500) moves over time. In general economic time series share a number of characteristics and behave in a surprisingly similar manner. This means that often one statistical methodology can be used to model almost all economic time series, usually quite successfully.

We can think of W_t as having three components: 1) a trend T_t, reflecting economic growth, 2) a cycle Y_t, reflecting the say the business cycle, and 3) a seasonal component S_t, reflecting such phenomena as the Christmas effect on consumption or the effect of weather on construction. We thus write:

$$W_t = f(T_t, Y_t, S_t). \tag{1.1}$$

The trend T_t and seasonality S_t will be non-stationary; that is the laws which determine how they move change over time. The cycle Y_t, on the other hand, will be assumed to be stationary; that is the laws governing how Y_t moves will not change over time.

To simplify the discussion, let us first assume that seasonality S_t plays no role. This is a reasonable assumption if we are working with seasonally adjusted data or data without a strong seasonal component such as exchange rates. Given this assumption we can ignore S_t and write:

$$W_t = f(T_t, Y_t). \tag{1.2}$$

A economically sensible functional form for $f\,(T_t, Y_t)$ is log-linear where:

$$W_t = f(T_t, Y_t) = T_t e^{Y_t}. \tag{1.3}$$

The cycle Y_t then determines if W_t is above or below trend. Thus if $Y_t > 0$ then $W_t > T_t$ and the series is above trend, while if $Y_t = 0$ then $W_t = T_t$ and the series is on trend while if $Y_t < 0$ then $W_t < T_t$ and the series is below trend.

In logarithms (1.3) becomes:

$$\ln(W_t) = \ln(T_t) + Y_t \tag{1.4}$$

so that $\ln(W_t)$ can be additively decomposed into a trend and cycle term.

1.2 The Trend

There are two ways of modelling the trend: the Trend Stationary or TS approach and the Difference Stationary or DS approach. We will first discuss the older and more traditional TS approach.

1.2.1 Trend Stationary Models

For the TS approach we imagine the trend grows deterministically as:

$$T_t = Ae^{\mu t} \tag{1.5}$$

where μ is the growth rate and A is the value of T_t at $t = 0$.

Substituting $T_t = Ae^{\mu t}$ into (1.3) we have:

$$W_t = Ae^{\mu t + Y_t}. \tag{1.6}$$

so if we then define X_t as the logarithm of W_t as:

$$X_t \equiv \ln(W_t) \tag{1.7}$$

then we get the linear relationship:

$$X_t = \alpha + \mu t + Y_t \tag{1.8}$$

where $\alpha = \ln(A)$.

Since Y_t the cycle will be assumed to be stationary, we see that X_t, is stationary except for the trend $\alpha + \mu t$, hence the terminology: trend stationary.

Suppose we have a sample of T observations: $\{W_1, W_2, \ldots W_T\}$ of a particular time series. For TS models we can obtain estimates of α and μ, the trend T_t and the cycle Y_t by applying ordinary least squares to (1.8). [1]

If the data are expressed as annual rates (i.e., GNP per year and not say GNP per quarter) then for the vast majority of economic time series we would expect values of μ roughly in the range:

$$0 \le \mu \le 0.1$$

[1] It can be shown under quite general conditions that OLS and GLS are asymptotically equivalent so that there is no loss of efficiency in using the simpler OLS.

reflecting growth between zero and ten percent per year.

For example with post-war quarterly U.S. real consumption, which is plotted below, we might obtain:

$$X_t = \underset{(1479.58)}{6.48} + \underset{(196.1)}{0.0084t} + Y_t \tag{1.9}$$

(the figures in brackets are t statistics). If the data are expressed as consumption per quarter, the coefficient on time is a quarterly growth rate. To convert to an annual growth rate we multiply by 4 to obtain:

$$0.0084 \times 4 = 0.034 \tag{1.10}$$

or an annual growth rate of about 3.4%. Using the rule of 72 we would expect this series to double about every:

$$\frac{72}{3.4} \approx 20 \text{ years.} \tag{1.11}$$

From the regression in (1.9)we can obtain an estimate of the cycle Y_t as the least squares residual or:

$$Y_t = X_t - 6.48 - 0.0084t. \tag{1.12}$$

To interpret the numerical value of Y_t use the following principle:

Definition 1 *If W_1 and W_2 are two numbers and $X_1 = \ln(W_1)$ and $X_2 = \ln(W_2)$ then*

$$\Delta X = X_2 - X_1$$

is (sensibly) defined as the percentage change from W_1 to W_2.

Proof. The usual definition of a percentage change is

$$g = \frac{W_2 - W_1}{W_1}.$$

From a first-order Taylor series

$$\ln(1 + x) \approx x$$

for x small (for example $\ln(1 + 0.04) = 0.0392 \approx 0.04$) so that:

$$\Delta X = \ln(W_2) - \ln(W_1) = \ln(1 + \frac{W_2 - W_1}{W_1}) \approx \frac{W_2 - W_1}{W_1} = g.$$

■

Remark 2 *Defining percentages for discrete changes always involves an arbitrary choice of the base. For example g uses W_1 as the base but we could equally well use W_2 or $\frac{W_1+W_2}{2}$ as the base.*

Remark 3 *Given g we can calculate ΔX from $\Delta X = e^{g} - 1$ while given ΔX we can calculate g from $g = \ln(1 + \Delta X)$. In practice there is usually little difference. For example if $g = 0.04$ or there is a 4% change, then $\Delta X = e^{0.04} - 1 = 0.0408$ which would indicate a 4.08% change.*

Using this definition of percentage we have:

Theorem 4 *The numerical value of Y_t is the percent by which the original series W_t is above or below trend: T_t.*

Proof. We have:

$$Y_t = \ln(W_t) - \ln(T_t) \tag{1.13}$$

so that Y_t is the percentage deviation of W_t from trend T_t. ■

For typical macroeconomic time series we might therefore expect magnitudes of Y_t in the range $-0.1 \leq Y_t \leq 0.1$, that is 10% above or below trend. You certainly would not expect say $Y_t = 3$, which would indicate the economy being 300% above trend. Such values (which do occur in practice!) typically indicate programming or data entry errors.

Example 5 *If you inspect Y_t for say U.S. consumer durables, you will see that in the early 1970's $Y_t \approx 0.05$ indicating that W_t was about 5% above trend while in the early 1980's $Y_t \approx -0.05$ indicating that W_t was about 5% below trend.*

1.2.2 Difference Stationary Models

An alternative approach, made popular in the 1970's by the work of Box and Jenkins, is the difference stationary (or DS) approach. Here we set

$$T_t = e^{\mu} W_{t-1}, \tag{1.14}$$

that is as the previous period's value of W_t increased by $\mu \times 100\%$ to reflect growth from period $t-1$ to t. Since W_{t-1} is random, it follows that the trend T_t is random. This is unlike the TS approach where the trend is nonrandom.. It is for this reason that people sometimes refer to DS models as *stochastic trend* models.

Substituting (1.14) into (1.3) we obtain:

$$W_t = W_{t-1} e^{\mu + Y_t}. \tag{1.15}$$

Again if we define $X_t \equiv \ln(W_t)$ we have:

$$X_t = X_{t-1} + \mu + Y_t \tag{1.16}$$

or

$$\Delta X_t = \mu + Y_t. \tag{1.17}$$

where we use Δ to denote differences so that:

$$\Delta X_t \equiv X_t - X_{t-1}. \tag{1.18}$$

Since $Y_t + \mu$ will turn out to be stationary, we see that X_t is stationary once it is differenced, hence the terminology: difference stationary.

Now to obtain an estimate of the cycle Y_t we regress ΔX_t on a constant. We might for example obtain:

$$\Delta X_t = \underset{(13.2)}{0.008} + Y_t \tag{1.19}$$

(where the figure in brackets is the t statistic) and the implied annual growth rate:

$$0.008 \times 4 = 0.032 \tag{1.20}$$

or 3.2% per year.

The cycle Y_t for the DS model can be obtained as the least squares residual from this regression or:

$$Y_t = \Delta X_t - 0.008. \tag{1.21}$$

In general the DS cycle will be very different than the TS cycle for the same series, displaying much less persistence.

To interpret the numerical value of Y_t we can use:

Theorem 6 *The numerical value of Y_t for the DS model is the amount by which the growth rate of the series W_t exceeds the average growth rate μ.*

Proof. This follows since the growth rate of W_t at time t is:

$$\Delta X_t = \ln(W_t) - \ln(W_{t-1}), \tag{1.22}$$

so that

$$Y_t = \Delta X_t - \mu \tag{1.23}$$

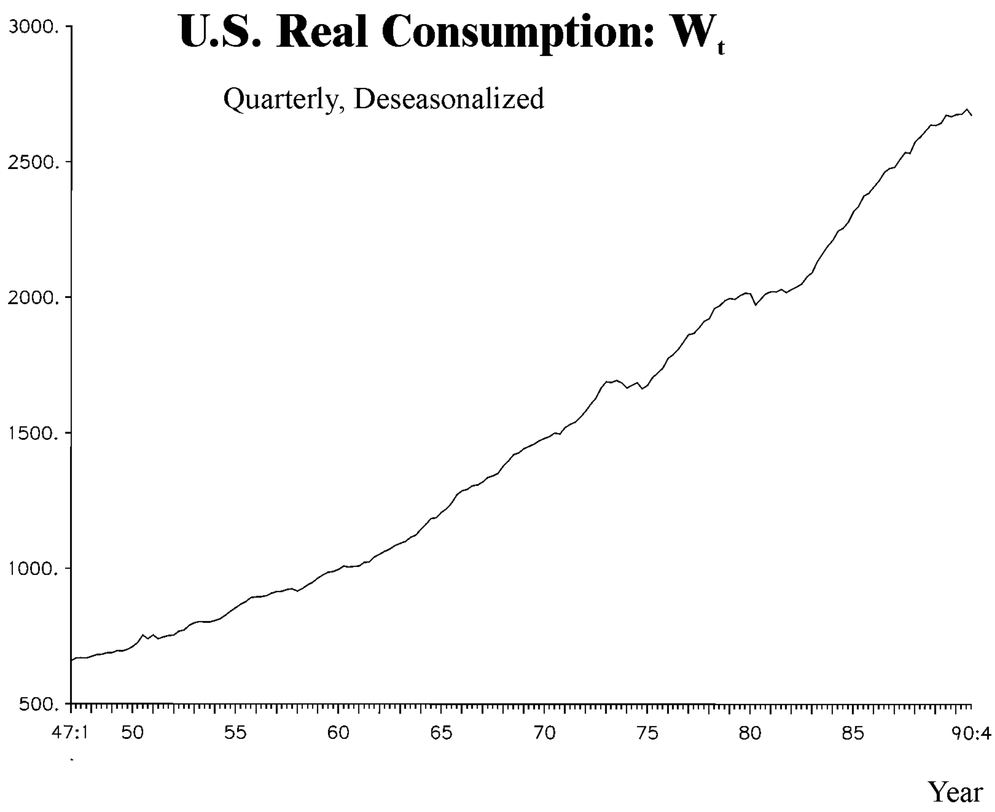
and hence Y_t is the extent to which the current growth rate ΔX_t is above or below the average growth rate μ. ■

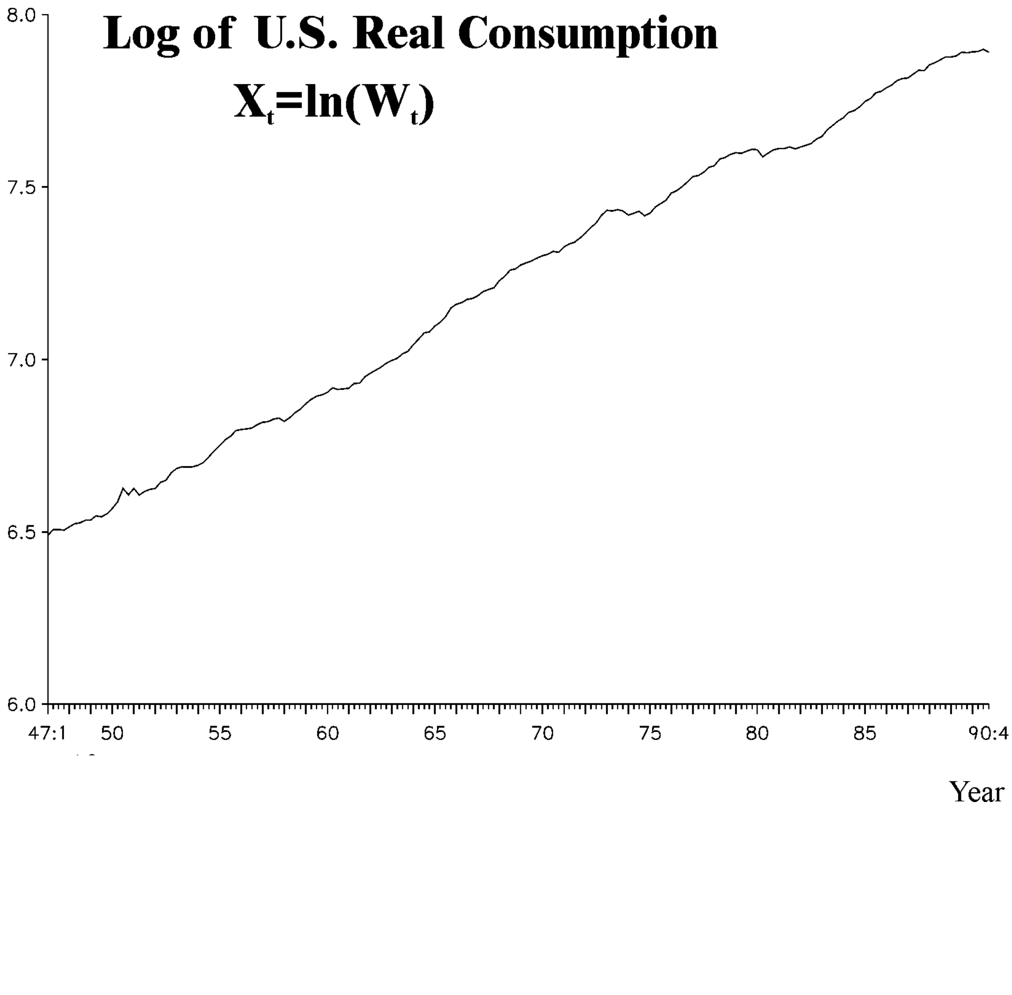
In general we would expect values of Y_t (in terms of annual growth rates) roughly in the range

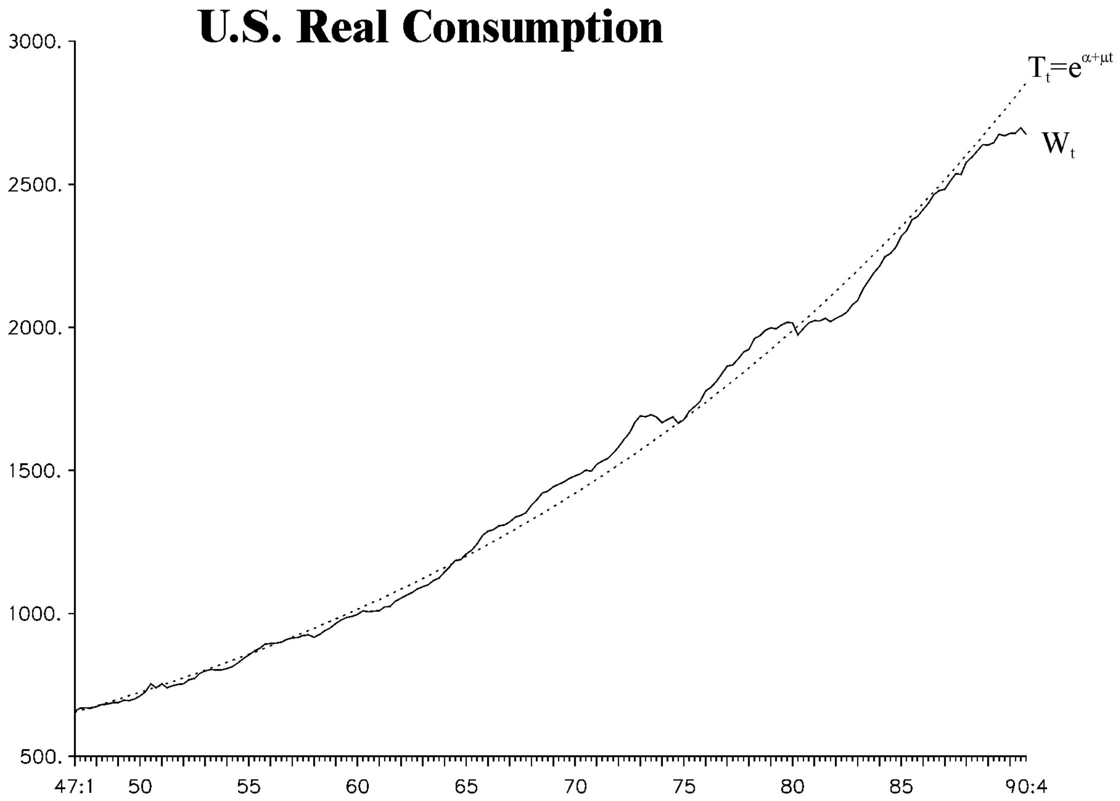
$$-0.03 \leq Y_t \leq 0.07. \tag{1.24}$$

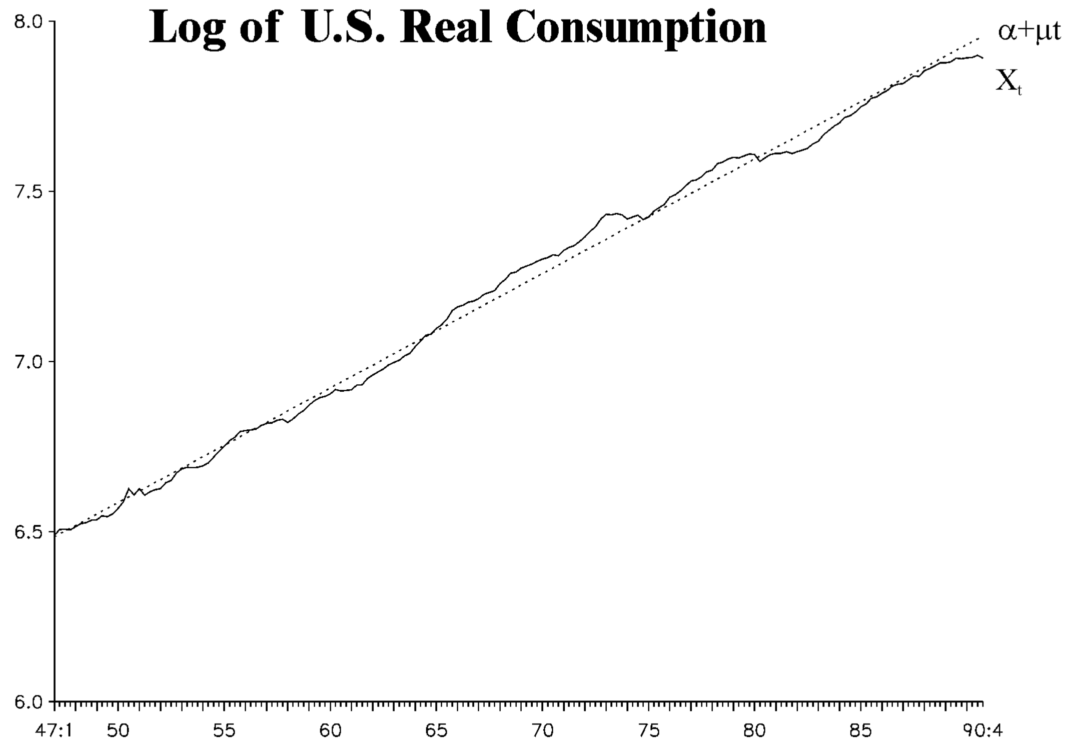
It is of course possible to find values outside this range but you should stop and think if you do. If you find say $Y_t = 10$, which would imply 1000% growth, then this clearly makes no economic sense so that something like a programming error or an incorrect data entry has occurred.

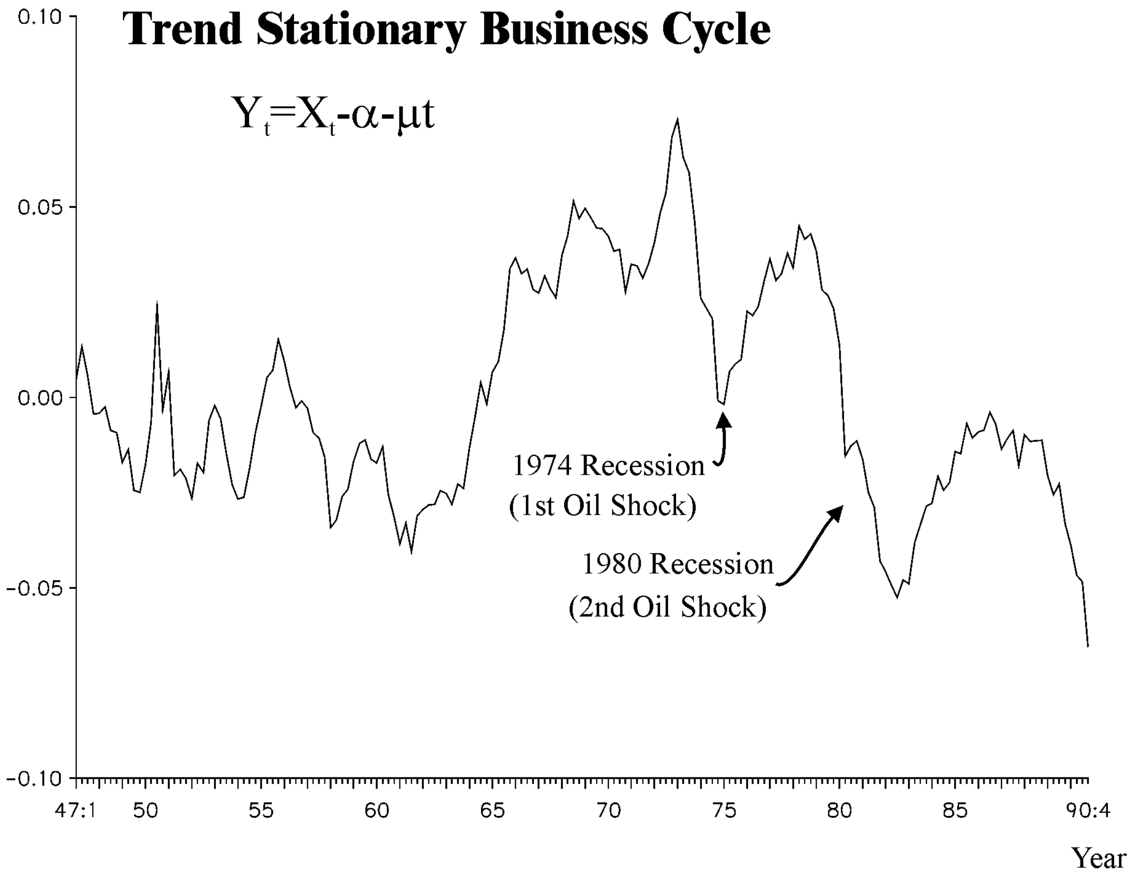
You can see the DS cycle Y_t for US real consumption, along with its TS counterpart in the graph above.

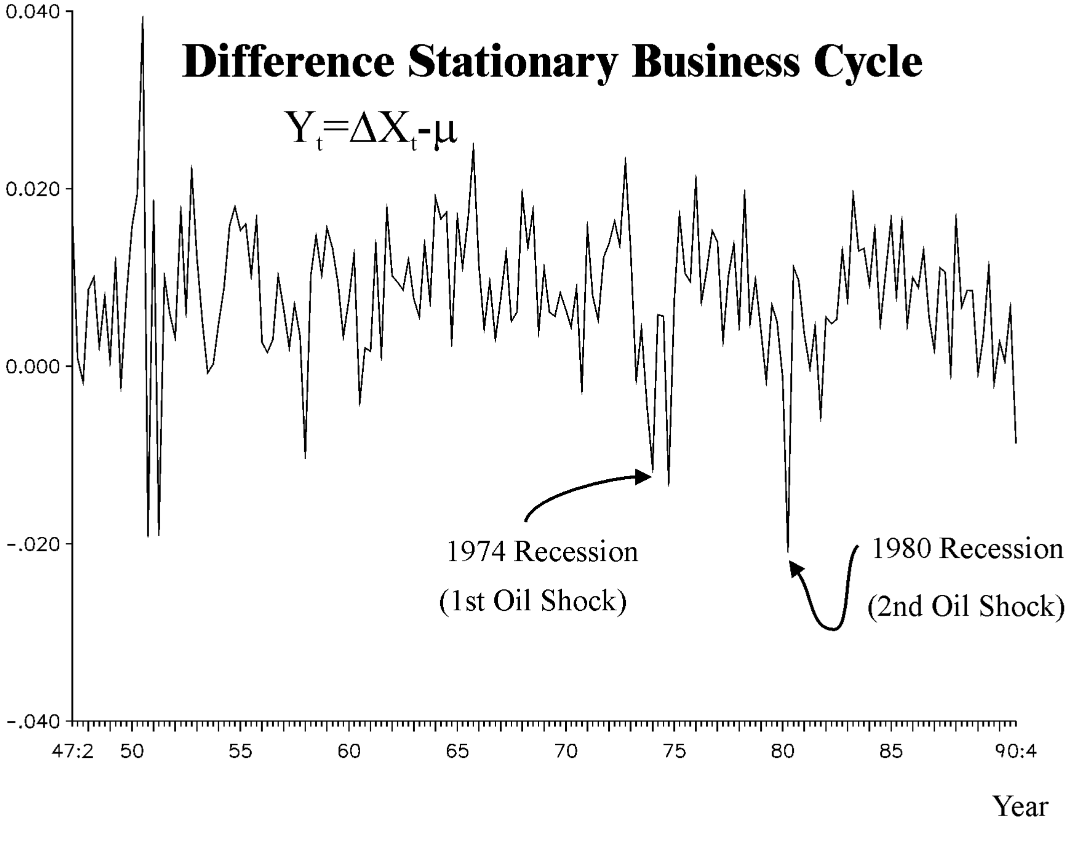

U.S. Real Consumption: W_t
Quarterly, Deseasonalized
3000.
2500.
2000.
1500.
1000.
500.
47:1
50
55
60
65
70
75
80
85
90:4
Year


Log of U.S. Real Consumption
$X_t=\ln(W_t)$
8.0
7.5
7.0
6.5
6.0
47:1
50
55
60
65
70
75
80
85
90:4
Year


U.S. Real Consumption


Log of U.S. Real Consumption


Trend Stationary Business Cycle
$Y_t = X_t - \alpha - \mu t$
0.10
0.05
0.00
-0.05
-0.10
47:1
50
55
60
65
70
75
80
85
90:4
Year
1974 Recession
(1st Oil Shock)
1980 Recession
(2nd Oil Shock)


Difference Stationary Business Cycle
$Y_t = \Delta X_t - \mu$
0.040
0.020
0.000
-.020
-.040
47:2
50
55
60
65
70
75
80
85
90:4
Year
1974 Recession
(1st Oil Shock)
1980 Recession
(2nd Oil Shock)

1.3 Seasonality

1.3.1 Introduction

Many economic time series have a strong seasonal component. For much of what we do, if you ignore this seasonality you will seriously compromise your results. This is because seasonality is a form of nonstationarity and generally speaking most of the methods we use require stationarity. So if after plotting a time series you find that it has a strong seasonal component, it is very important that you either 1) obtain the seasonally adjusted version of the series or 2) seasonally adjust the data yourself. Below we discuss a few methods for doing the seasonal adjustment yourself.

1.3.2 TS and Seasonal Dummies

Suppose that we are using the TS approach. Without seasonality we would have:

$$X_t = \alpha + \mu t + Y_t. \tag{1.25}$$

Suppose there are S is the number of periods in one year. For quarterly data then $S = 4$ while for monthly data $S = 12$. One strategy for dealing with seasonality is to replace the intercept α with S dummy variables

$$d_{1t}, d_{2t}, \ldots d_{st}$$

so that:

$$X_t = \sum_{j=1}^{S} \alpha_j d_{jt} + \mu t + Y_t. \tag{1.26}$$

To obtain Y_t, the cycle, would then run least squares on X_t with S seasonal dummies and a time trend and take Y_t as the least squares residual.

For example with quarterly data (or $S = 4$) one would have four dummy variables so that:

$$X_t = \alpha_1 d_{1t} + \alpha_2 d_{2t} + \alpha_3 d_{3t} + \alpha_4 d_{4t} + \mu t + Y_t \tag{1.27}$$

where:

$$\begin{aligned} d_{it} &= 1, \text{ if } t \text{ is in quarter } i, \ d_{it} = 0 \text{ otherwise} \\ i &= 1, 2, 3, 4. \end{aligned} \tag{1.28}$$

With consumption data we would expect $\alpha_4 > \alpha_1$, reflecting the Christmas effect on consumption.

1.3.3 DS with Seasonal Dummies

For the DS model with seasonality a reasonable assumption is that the growth rate varies according to the period we are in. Thus instead of:

$$\Delta X_t = \mu + Y_t \tag{1.29}$$

we would have:

$$\Delta X_t = \sum_{j=1}^{S} \mu_j d_{jt} + Y_t. \tag{1.30}$$

To obtain the cycle Y_t one would therefore regress ΔX_t on the S seasonal dummies and obtain Y_t as the least squares residual.

For example with quarterly data we would have:

$$\Delta X_t = \mu_1 d_{1t} + \mu_2 d_{2t} + \mu_3 d_{3t} + \mu_4 d_{4t} + Y_t \tag{1.31}$$

so that to obtain the cycle Y_t one would regress ΔX_t on four seasonal dummies and obtain Y_t as the least squares residual.

1.3.4 Seasonal Differencing

Another approach to seasonality which was made popular by Box and Jenkins is to seasonally difference. Here instead of (1.14) the trend takes the form:

$$T_t = W_{t-s} e^{\mu t}. \tag{1.32}$$

This leads to:

$$X_t - X_{t-s} = \mu + Y_t \tag{1.33}$$

so that instead of differencing 1 period as we normally do for DS models, we instead difference say S periods.

For example with quarterly data instead of regressing:

$$X_t - X_{t-1} = \mu + Y_t$$

we would run the regression:

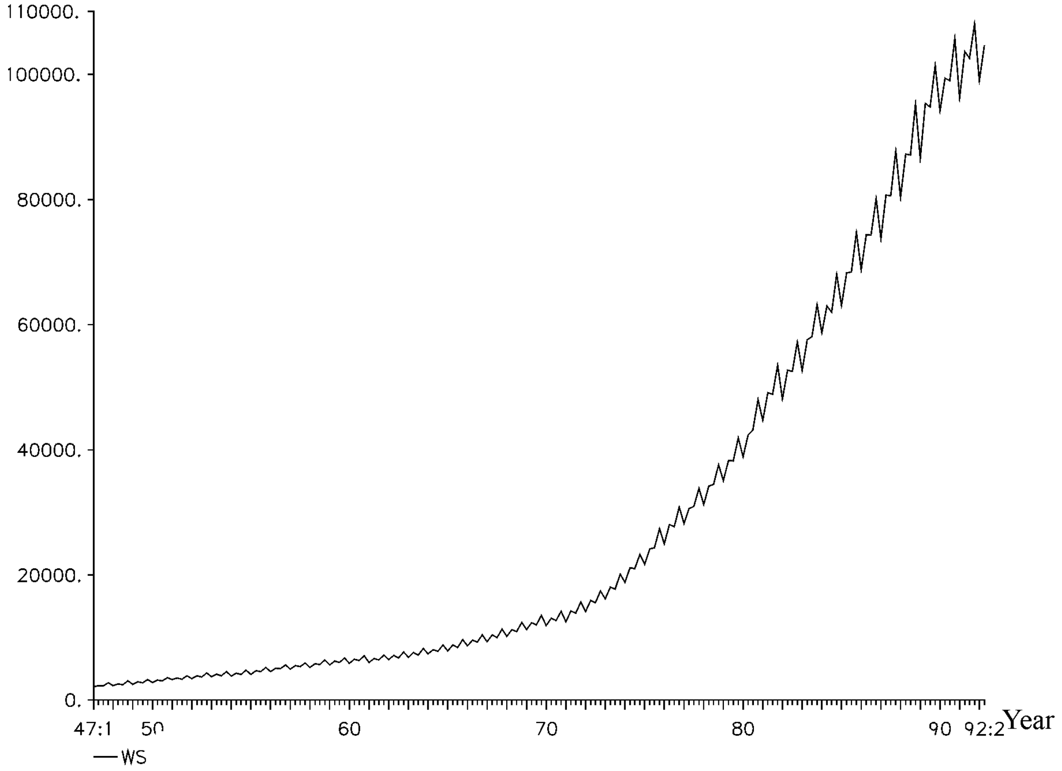
$$X_t - X_{t-4} = \mu + Y_t$$

and obtain Y_t as the least squares residual.

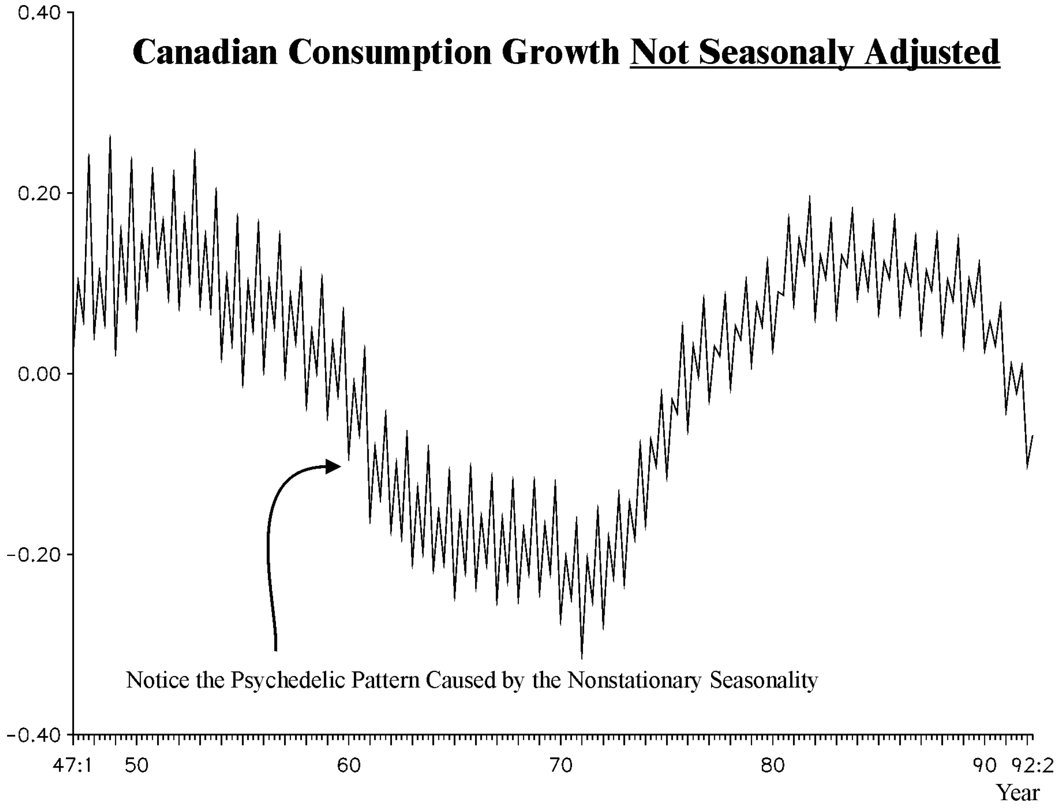
1.4 Modeling the Cycle

The trend usually accounts for most of the movement of W_t, but it is the cycle Y_t that is the more difficult and more interesting to model. The next three chapters will deal with modelling Y_t and the theory behind these models.

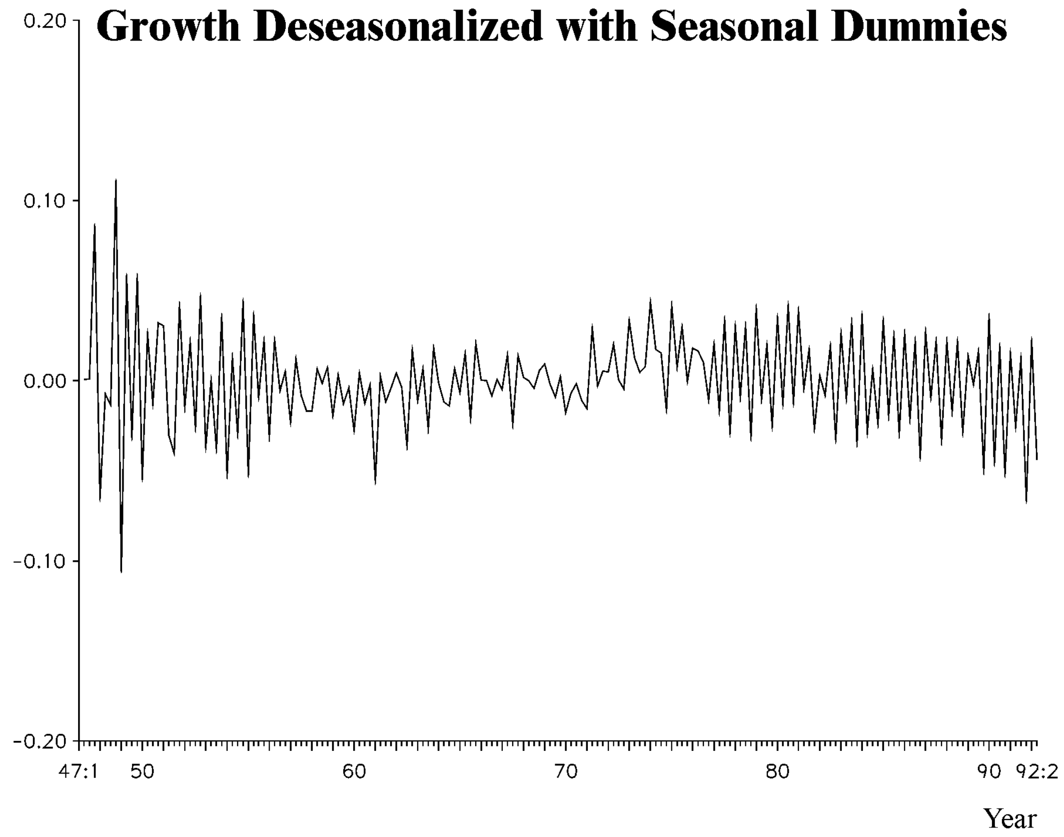
Canadian Real Consumption Not Seasonaly Adjusted



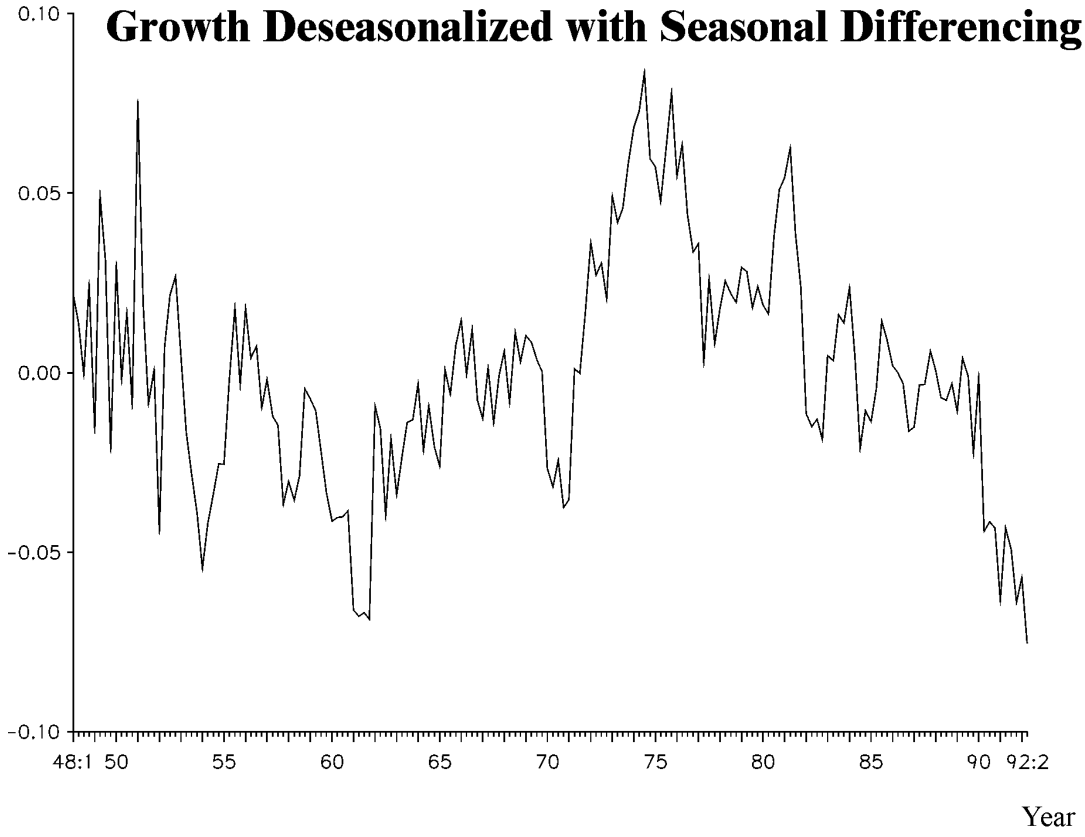
Canadian Consumption Growth Not Seasonaly Adjusted



Notice the Psychedelic Pattern Caused by the Nonstationary Seasonality



Growth Deseasonalized with Seasonal Dummies



Growth Deseasonalized with Seasonal Differencing

The approach we use in time series analysis is to think of Y_t as a stationary random (or stochastic) process such as a first-order autoregressive process or AR(1) where:

$$Y_t = \phi Y_{t-1} + a_t \tag{1.34}$$

or alternatively a first-order moving average process or MA(1) where:

$$Y_t = a_t + \theta a_{t-1}. \tag{1.35}$$

More generally we will be interested in autoregressive moving average processes with p lags of Y_t and q lags of a_t where:

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + a_t + \sum_{j=1}^{q} \theta_j a_{t-j}.$$

This is referred to as an ARMA(p,q) process.

Chapter 2

Stationary Stochastic Processes

2.1 Introduction

Stationarity is one of the fundamental concepts in time series analysis. Roughly speaking a random process Y_t is stationarity if the probabilistic laws determining Y_t do not change over time. Quite a bit of mathematical apparatus is necessary to express this idea precisely, but fortunately this will not be necessary for our purposes.

Neither growth nor seasonality are stationary. For example the growth of GNP means that GNP is likely to be much higher in 1999 than it was in 1947 and so the laws governing GNP in 1999 are not the same as the laws of 1947.

Similarly seasonality for consumption implies that consumption during the Christmas quarter is likely to be higher than consumption in the January-March quarter and so the laws governing consumption differ in the two quarters.

It is, however, reasonable to assume that growth rates of many economic time series are stationary; for example that the probability of a recession (say defined as two consecutive quarters of negative growth rate in GNP) was the same in 1963 as it is in 1999. Once an economic time series has been detrended and seasonally adjusted, stationarity is generally a natural assumption to make.

There are in fact many different mathematical definitions of stationarity. For our purposes second order or weak stationarity is usually sufficient and can be easily stated. It says that a time series Y_t is stationary if all means, variances and covariances are the same for all time periods. Thus:

Definition 7 ***Stationarity:*** *A time series Y_t is stationary if for all periods t, s and k :*

$$\begin{aligned} E\left[Y_t\right] &= E\left[Y_s\right] \qquad (2.1) \\ Var\left[Y_t\right] &= Var\left[Y_s\right] \\ Cov\left[Y_t, Y_{t-k}\right] &= Cov\left[Y_s, Y_{s-k}\right]. \end{aligned}$$

One of the reasons this definition is sufficient is that most of the time series we will be dealing with are normally distributed or Gaussian processes defined as follows:

Definition 8 ***Gaussian Process:*** *A random process Y_t is Gaussian if for any s periods $t_1, t_2, \ldots t_s$ the $s \times 1$ random vector $\tilde{Y} = [Y_{t_i}]$ has a multivariate normal distribution or:*

$$\tilde{Y} \sim N[\mu, \Sigma].$$

or the density of $\tilde{Y}$ is given by:

$$p\left(\tilde{Y}\right) = (2\pi)^{-\frac{s}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\tilde{Y} - \mu\right)^T \Sigma^{-1}\left(\tilde{Y} - \mu\right)\right).$$

Since the normal distribution only depends on the mean in μ and the variances and covariances in Σ, to show stationarity one only needs that μ and Σ are the same for all periods.

The mean of Y_t is not very important or interesting for stationary time series. Many of the derivations that we will do are greatly simplified if we assume that Y_t has a mean of zero. Given stationarity this can be done without any loss of generality since if Y_t does not have a mean of zero but say μ, then it is always possible to construct another time series $Y_t - \mu$ that does have a mean of zero. This is a standard assumption in the time series literature and one you need to get used to.

We state this formally as follows:

Theorem 9 *If Y_t^* is any stationary time series with a mean of μ then if $Y_t = Y_t^* - \mu$ then: $E[Y_t] = 0$.*

Remark 10 *This turns out to be a natural assumption when using either the TS or DS approach since Y_t is in either case a least-squares residual from a regression with a constant term and consequently must have a sample mean of* 0.

The assumption that $E[Y_t] = 0$ simplifies derivations because variances and covariances can be expressed as expectations of products as:[1]

$$\begin{aligned} Var[Y_t] &= E\left[Y_t^2\right] \\ Cov[Y_t, Y_s] &= E[Y_t Y_s]. \end{aligned} \tag{2.2}$$

[1] Since if $E[X] = 0$ and $E[Y] = 0$ then:

$$\begin{aligned} Var[X] &= E\left[X^2\right] - E[X]^2 \\ &= E\left[X^2\right] \end{aligned}$$

and

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= E[XY] \end{aligned}$$

This leads to an important principle:

Proposition 11 ***In general when dealing with stationary random variables with a mean of zero, expectations of squares are variances and expectations of products are covariances.***

2.2 The Short-Memory Property

Many of the models that we will be considering will have the property that they quickly forget or quickly become independent of what occurs either in the distant future or the distant past. This forgetting occurs at an exponential rate which represents a very rapid type of decay.

For example if you have a pie in the fridge and you eat one-half of the pie each day, you will quickly have almost no pie. After only ten days you would have:

$$\left(\frac{1}{2}\right)^{10} = \frac{1}{1024}$$

or about one-thousandth of a pie; maybe a couple of crumbs.

We will see that for stationary ARMA(p,q) processes, the infinite moving average weights: ψ_k, the autocorrelation function $\rho(k)$ and the forecast function $E_t[Y_{t+k}]$, all functions of the number of periods k, all have the short-memory property which we now define:

Definition 12 ***Short-Memory:*** *Let P_k for $k = 0, 1, 2, \ldots \infty$ be some numerical property of a stationary time series which depends on k, the number of periods. We say P_k displays a short-memory or $P_k = O\left(\tau^k\right)$ if*

$$|P_k| \leq A\tau^k$$

where $A \geq 0$ and $0 < \tau < 1$.

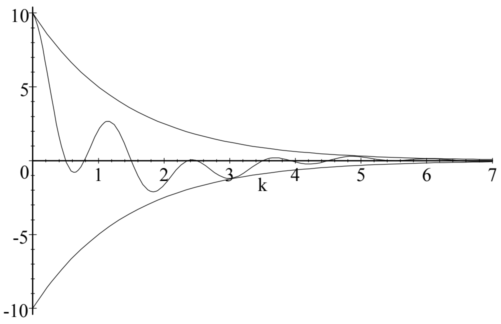
If $P_k = O\left(\tau^k\right)$ or if P_k has a short-memory then P_k decays rapidly in the same, manner that is at least as fast as τ^k decays to zero as $k \to \infty$. For example if:

$$P_k = 10\cos(2k)\left(-\frac{1}{2}\right)^k \tag{2.3}$$

then P_k decays rapidly in a manner which is bounded by exponential decay since $|\cos(2k)| \leq 1$ and so we have:

$$|P_k| \leq 10\left(\frac{1}{2}\right)^k = A\tau^k \tag{2.4}$$

where $\tau = \frac{1}{2}$ and $A = 10$. This is illustrated in the plot below:



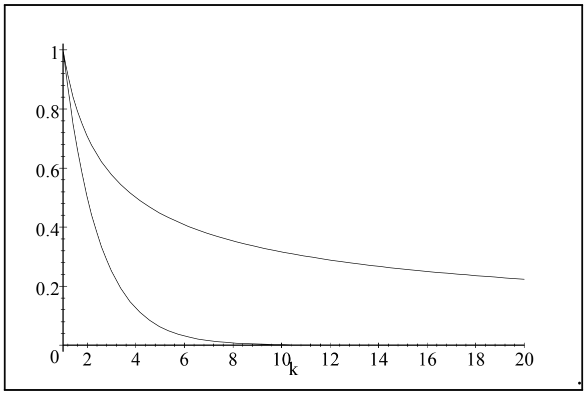
$$P_k = 10\cos(2k)\left(-\tfrac{1}{2}\right)^k$$

Not everything decays so rapidly. For example if we reverse the $\frac{1}{2}$ and the k in $\left(\frac{1}{2}\right)^k$ we obtain:

$$Q_k = \frac{1}{k^{\frac{1}{2}}} = k^{-\frac{1}{2}}. \tag{2.5}$$

This is *hyperbolic* decay, a much slower rate of decay.

We can compare the hyperbolic decay of $Q_k = k^{-\frac{1}{2}}$ with the exponential decay of $R_k = \left(\frac{1}{2}\right)^{k-1}$. For these two example both are equal for $k = 1$ but R_k decays much faster than Q_k as can be seen by the diagram below:



Plot of Q_k and R_k

Note that R_k is effectively equal to zero for $k \geq 8$ while Q_k is still large for $k = 20$.

Let us therefore define long-memory as being bounded by hyperbolic decay as follows:

Definition 13 ***Long-Memory:*** *Let P_k for $k = 0, 1, 2, \ldots \infty$ be some numerical property of a stationary time series which depends on k, the number of periods. We say P_k displays a long-memory or $P_k = O\left(k^{-\tau}\right)$ if*

$$|P_k| \leq Ak^{-\tau}$$

where $A \geq 0$ and $\tau > 0$.

2.3 The AR(1) Model

The simplest interesting model for Y_t is a first-order autoregressive process or AR(1) which can be written as:

$$Y_t = \phi Y_{t-1} + a_t, \quad a_t \sim i.i.n(0, \sigma^2) \ , \tag{2.6}$$

where $i.i.n.(0, \sigma^2)$ means that a_t is independently and identically distributed ($i.i.d.$) with a normal distribution with mean 0 and variance σ^2 so that the density of a_t is:

$$p\left(a_t\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}a_t^2}. \tag{2.7}$$

We can attempt to calculate $E\left[Y_t\right]$ by taking expectations of both sides of (2.6) to obtain:

$$\begin{aligned} E\left[Y_t\right] &= \phi E[Y_{t-1}] + E[a_t] \\ &= \phi E[Y_{t-1}] \end{aligned} \tag{2.8}$$

since $E[a_t] = 0$. We now need to find $E\left[Y_{t-1}\right]$. We could try the same approach with $E\left[Y_{t-1}\right]$ since $Y_{t-1} = \phi Y_{t-2} + a_{t-1}$ from which we would conclude that: $E\left[Y_{t-1}\right] = \phi E\left[Y_{t-2}\right]$ so that:

$$E\left[Y_t\right] = \phi^2 E\left[Y_{t-2}\right]; \tag{2.9}$$

but now we now need to find $E\left[Y_{t-2}\right]$. Clearly this process will never end.

If, however, we assume stationarity then it is possible to break this infinite regress since by the definition of stationarity in Definition 7:

$$E[Y_t] = E[Y_{t-1}]. \tag{2.10}$$

It then follows from (2.8) that:

$$E[Y_t] = \phi E\left[Y_t\right] \tag{2.11}$$

or

$$\left(1 - \phi\right) E\left[Y_t\right] = 0. \tag{2.12}$$

Assuming that $\phi \neq 1$ (which turns out to be necessary for stationarity) we conclude that:

Theorem 14 *If Y_t is an AR(1) given by* (2.6) *and Y_t is stationary then:*

$$E\left[Y_t\right] = 0.$$

The same trick can be use to calculate $Var[Y_t]$ for an AR(1). From (2.6) we have:

$$\begin{aligned} Var[Y_t] &= Var\left[\phi Y_{t-1} + a_t\right] \qquad (2.13) \\ &= \phi^2 \underbrace{Var[Y_{t-1}]}_{=Var[Y_t]} + 2\phi \underbrace{Cov[Y_{t-1}, a_t]}_{=0} + \underbrace{Var[a_t]}_{=\sigma^2} \end{aligned}$$

since by stationarity:

$$Var[Y_t] = Var[Y_{t-1}] \qquad (2.14)$$

and a_t is *i.i.d.* and hence uncorrelated with Y_{t-1}. Solving for $Var[Y_t]$ we obtain:

Theorem 15 *If Y_t is an AR(1) given by* (2.6) *and Y_t is stationary then:*

$$Var[Y_t] = \frac{\sigma^2}{1-\phi^2}. \qquad (2.15)$$

The formula (2.15) only makes sense if the variance is non-negative and finite; that is if:

$$0 \leq Var[Y_t] < \infty. \qquad (2.16)$$

From this we conclude that a necessary condition for stationarity is that: $-1 < \phi < 1$ since if say $\phi > 1$ the variance would be negative while if $\phi = \pm 1$ the variance would be infinite.

This leads us to :

Theorem 16 *Y_t is a stationary AR(1) process only if*

$$-1 < \phi < 1.$$

Given stationarity it can be shown that:

Theorem 17 *The unconditional distribution of a stationary AR(1) process is:*

$$Y_t \sim N\left[0, \frac{\sigma^2}{1-\phi^2}\right] \qquad (2.17)$$

for all t.

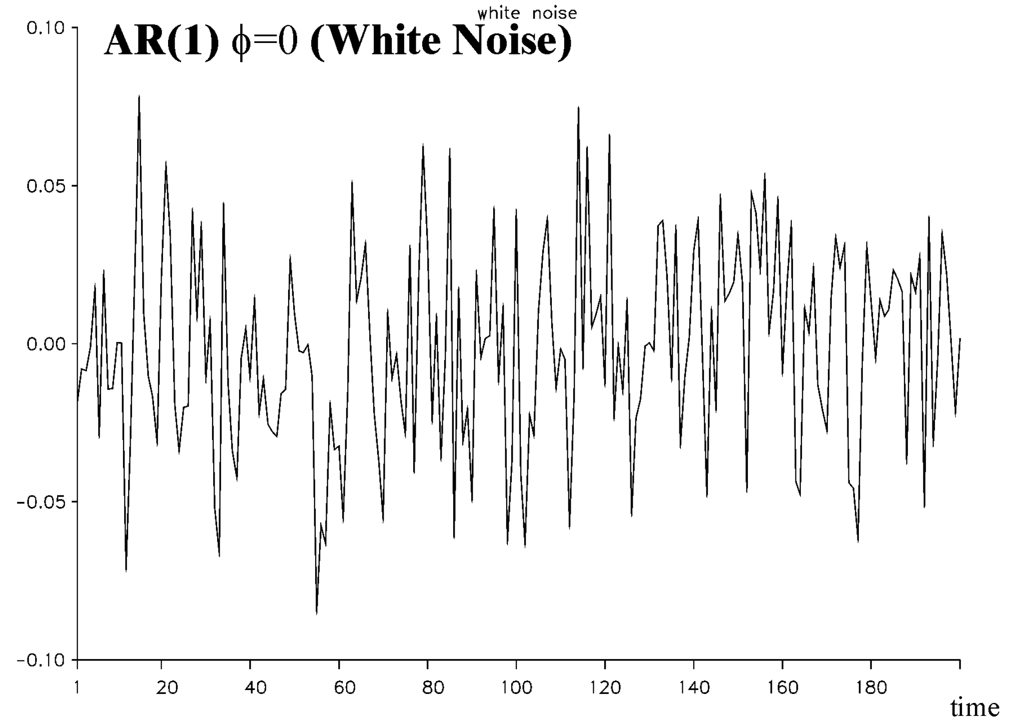
From this it then follows then that a band given by:

$$0 \pm 1.96 \frac{\sigma}{\sqrt{1-\phi^2}} \qquad (2.18)$$

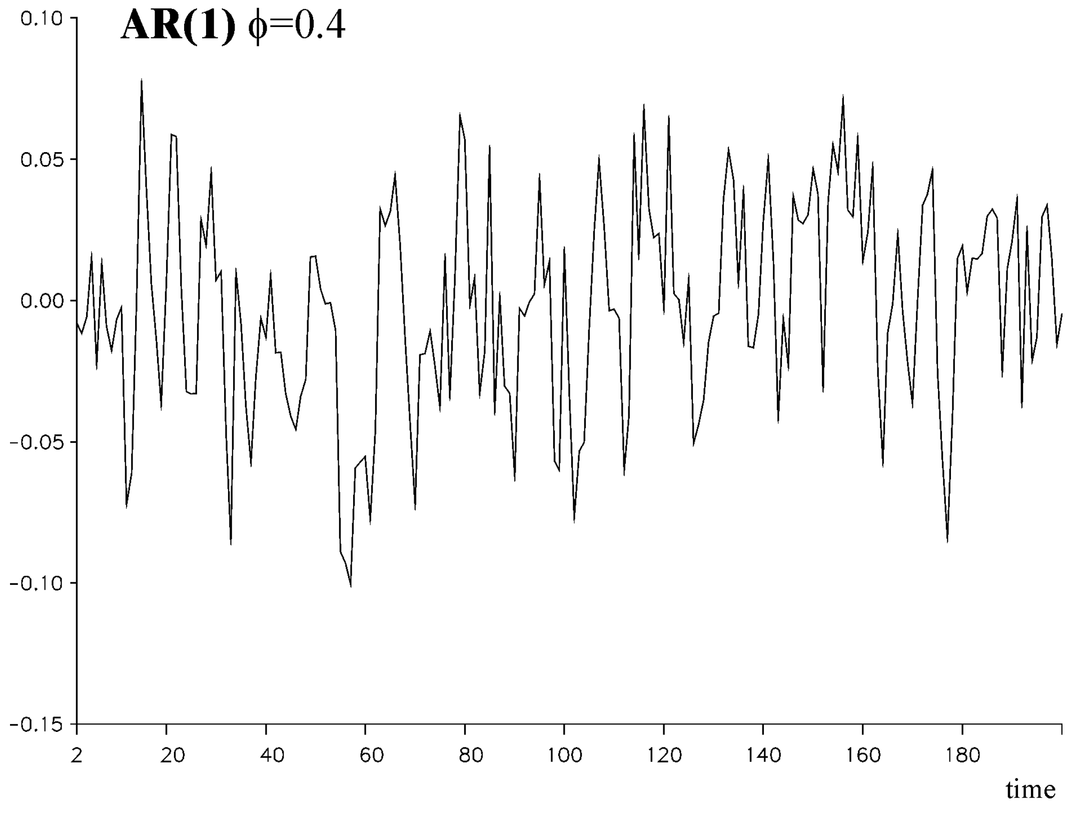
would contain 95% of all realizations of Y_t.

The covariances from an AR(1) process can be calculated recursively using the following result:

Simulated AR(1) Processes

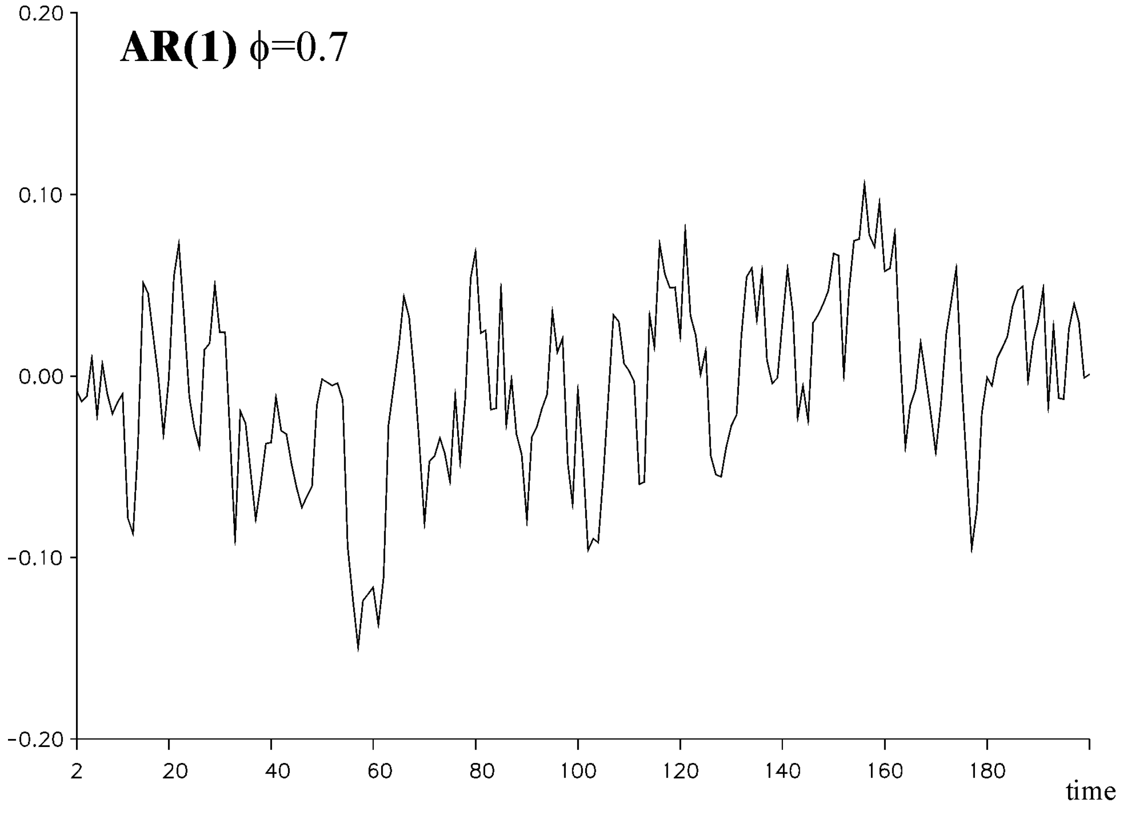


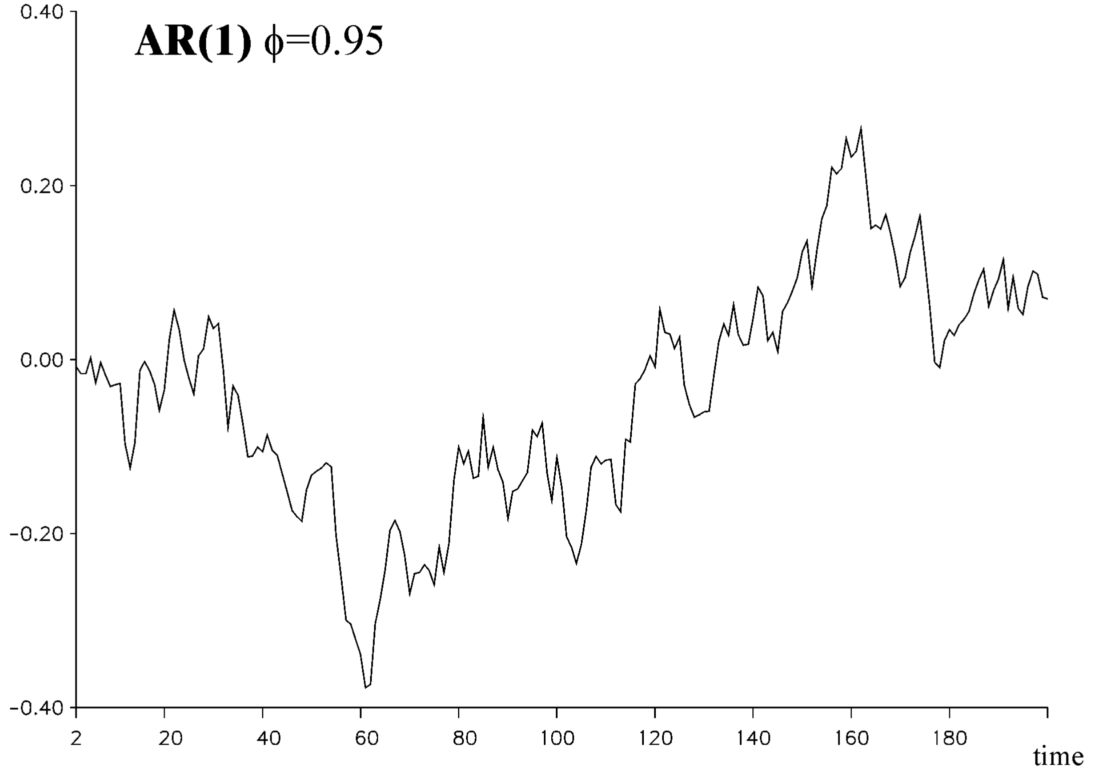
AR(1) ϕ=0 (White Noise)



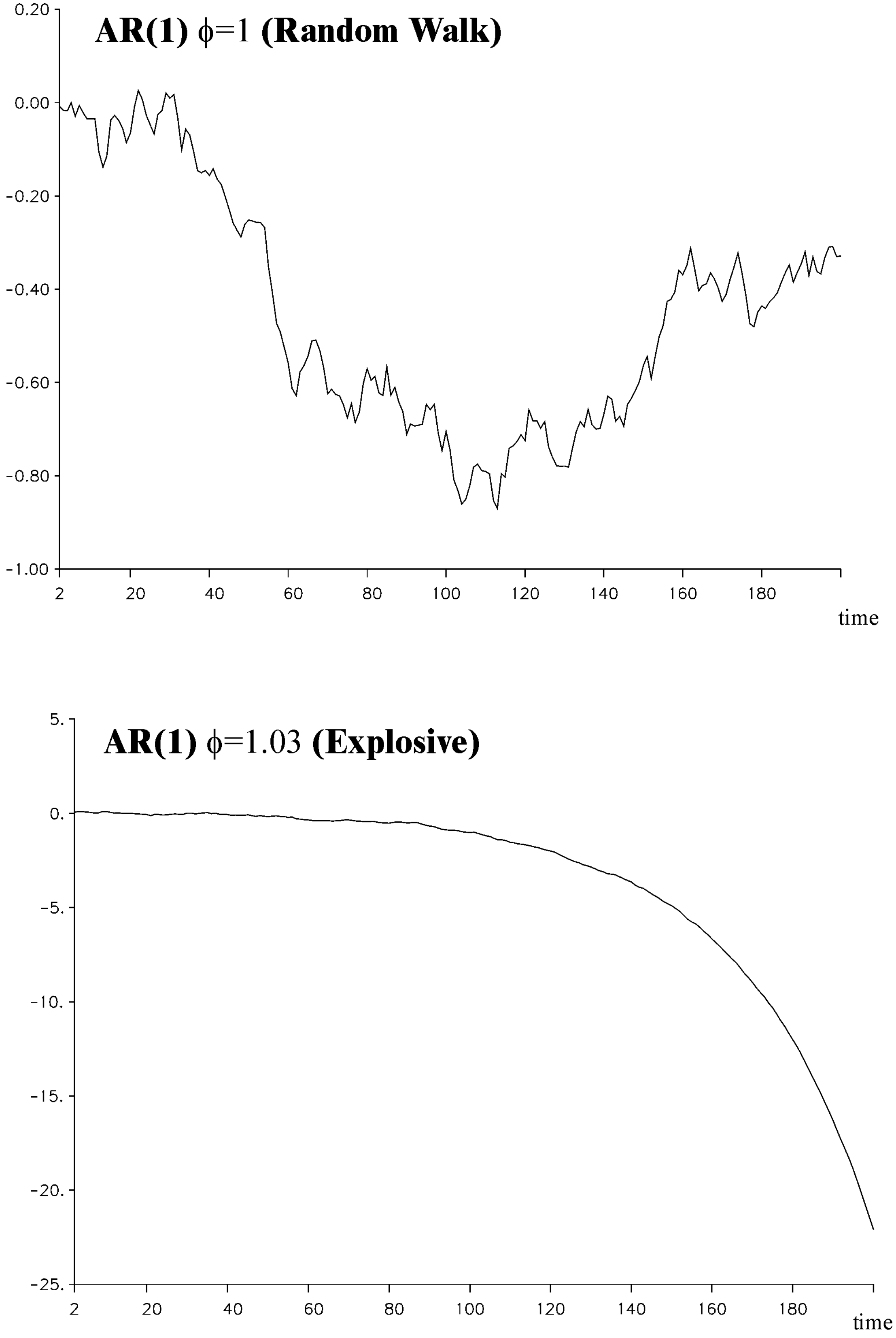
AR(1) ϕ=0.4

Simulated AR(1) Processes





Simulated AR(1) Processes



Theorem 18 *For an AR(1) with $k > 0$:*

$$Cov\left[Y_t, Y_{t-k}\right] = \phi Cov\left[Y_t, Y_{t-(k-1)}\right].$$

Proof. Using (2.6) we have:

$$\begin{aligned} Cov\left[Y_t, Y_{t-k}\right] &= E[Y_t Y_{t-k}] \\ &= E[(\phi Y_{t-1} + a_t)\, Y_{t-k}] \\ &= \phi E[Y_{t-1} Y_{t-k}] + E[a_t Y_{t-k}] \\ &= \phi Cov\left[Y_{t-1}, Y_{t-k}\right] \\ &= \phi Cov\left[Y_t, Y_{t-(k-1)}\right] \end{aligned}$$

where $E[a_t Y_{t-k}] = 0$ since a_t is uncorrelated with the past Y_{t-k} while:

$$\begin{aligned} Cov\left[Y_{t-1}, Y_{t-k}\right] &= Cov\left[Y_s, Y_{s-(k-1)}\right] \\ &= Cov\left[Y_t, Y_{t-(k-1)}\right] \end{aligned}$$

where $s = t - 1$ and using Definition 7. ■

We already know that:

$$Cov\left[Y_t, Y_t\right] = \frac{\sigma^2}{1-\phi^2}.$$

To calculate $Cov\left[Y_t, Y_{t-1}\right]$ we then use Theorem 18 to obtain:

$$\begin{aligned} Cov\left[Y_t, Y_{t-1}\right] &= \phi Cov\left[Y_t, Y_t\right] \\ &= \frac{\phi\sigma^2}{1-\phi^2}. \end{aligned}$$

Similarly:

$$\begin{aligned} Cov\left[Y_t, Y_{t-2}\right] &= \phi Cov\left[Y_t, Y_{t-1}\right] \\ &= \frac{\phi^2\sigma^2}{1-\phi^2}. \end{aligned}$$

More generally using Theorem 18 on $Cov\left[Y_t, Y_{t-(k-1)}\right]$ we obtain:

$$Cov\left[Y_t, Y_{t-(k-1)}\right] = \phi Cov\left[Y_t, Y_{t-(k-2)}\right] \tag{2.19}$$

so that repeating this argument we conclude that:

$$\begin{aligned} Cov\left[Y_t, Y_{t-k}\right] &= \phi^k Cov\left[Y_t, Y_t\right] && (2.20) \\ &= \frac{\phi^k\sigma^2}{1-\phi^2} && (2.21) \end{aligned}$$

and so we conclude that:

Theorem 19 *For a stationary AR(1) for $k \geq 0$ we have:*

$$Cov\left[Y_t, Y_{t-k}\right] = \phi^k \left(\frac{\sigma^2}{1-\phi^2}\right).$$

This then implies that $Cov\left[Y_t, Y_{t-k}\right] = O\left(\tau^k\right)$ or $Cov\left[Y_t, Y_{t-k}\right]$ has the short-memory property given in Section 2.2 since:

$$Cov\left[Y_t, Y_{t-k}\right] \leq A\tau^k \tag{2.22}$$

with $A = Var\left[Y_t\right]$ and $\tau = \phi$. This means that Y_t rapidly forgets its past history.

2.4 The Autocovariance Function

An important implication of stationarity is that the covariance between the business cycle in say the first and third quarters of say 1999 is the same as the covariance between the business cycle the first and third quarters of say 1963. In general covariances only depend on the number of periods separating Y_t and Y_s so that:

Theorem 20 *If Y_t is stationary then $Cov\left[Y_{t_1}, Y_{t_2}\right]$ depends only on $k = t_1 - t_2$; that is the number of periods separating t_1 and t_2.*

Since we will often be focusing on covariances, and since $Cov\left[Y_t, Y_{t-k}\right]$ only depends on k, let us define this as a function of k as: $\gamma\left(k\right)$, which we will refer to as the autocovariance function so that:

Definition 21 ***Autocovariance Function:*** *Let Y_t be a stationary time series with $E[Y_t] = 0$. The autocovariance function for Y_t, denoted as $\gamma(k)$, is defined for $k = 0, \pm 1, \pm 2, \pm 3, \ldots \pm \infty$ as:*

$$\gamma(k) \equiv E[Y_t Y_{t-k}] = Cov[Y_t, Y_{t-k}].$$

We have the following results for the autocovariance function:

Theorem 22 $\gamma(0) = Var[Y_t] > 0$

Theorem 23 $\gamma(k) = E[Y_t Y_{t-k}] = E[Y_s Y_{s-k}]$ *for any t and s.*

Theorem 24 $\gamma(-k) = \gamma(k)$ *($\gamma\left(k\right)$ is an even function)*

Theorem 25 *Let $t_1, t_2, \ldots t_k$ be any k periods then the symmetric $k \times k$ matrix Γ with i, j^{th} element given by $\Gamma_{ij} = \gamma\left(t_i - t_j\right)$ or:*

$$\Gamma = \begin{bmatrix} \gamma\left(0\right) & \gamma\left(t_1 - t_2\right) & \gamma\left(t_1 - t_3\right) & \cdots & \gamma\left(t_1 - t_k\right) \\ \gamma\left(t_1 - t_2\right) & \gamma\left(0\right) & \gamma\left(t_2 - t_3\right) & \cdots & \gamma\left(t_2 - t_k\right) \\ \gamma\left(t_1 - t_3\right) & \gamma\left(t_2 - t_3\right) & \gamma\left(0\right) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \gamma\left(t_{k-1} - t_k\right) \\ \gamma\left(t_1 - t_k\right) & \gamma\left(t_2 - t_k\right) & \cdots & \gamma\left(t_{k-1} - t_k\right) & \gamma\left(0\right) \end{bmatrix}$$

is positive semi-definite.

Theorem 26 *If*

$$Var\left[\sum_{i=1}^{k} a_i Y_{t_i}\right] > 0$$

whenever there is at least one $a_i \neq 0$ then Γ defined in Theorem 25 is positive definite.

Theorem 27 $|\gamma(k)| \leq \gamma(0)$.

Proof. Theorem 22 follows from the fact that the covariance between a random variable and itself is its variance. Theorem 23 follows by stationarity. To prove Theorem 24 note that:

$$\begin{aligned}\gamma(k) &= E[Y_t Y_{t-k}] \\ &= E[Y_{t-k} Y_t] \\ &= E[Y_s Y_{s-(-k)}] \text{ where } s = t-k. \\ &= E[Y_t Y_{t-(-k)}] \text{ (by stationarity)} \\ &= \gamma\left(-k\right).\end{aligned}$$

To prove Theorem 25 note that if

$$\tilde{Y} = \sum_{i=1}^{k} a_i Y_{t_i}$$

then:

$$Var\left[\tilde{Y}\right] = a^T \Gamma a \geq 0$$

where a is the $n \times 1$ vector of the a_i $'s$. It follows then that Γ is positive semi-definite. If $Var\left[\tilde{Y}\right] > 0$ for any $a \neq 0$ then:

$$a^T \Gamma a > 0$$

and so Γ is positive definite. This proves Theorem 26. Finally note that if $\tilde{Y} = aY_t + bY_{t-k}$ then

$$\begin{aligned}Var[aY_t + bY_{t-k}] &= a^2 \underbrace{Var[Y_t]}_{\gamma(0)} + 2ab\underbrace{Cov[Y_t, Y_{t-k}]}_{\gamma(k)} + b^2 \underbrace{Var[Y_{t-k}]}_{\gamma(0)} \\ &= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \gamma(0) & \gamma(k) \\ \gamma(k) & \gamma(0) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \geq 0\end{aligned}$$

so that Γ is given by:

$$\Gamma = \begin{bmatrix} \gamma(0) & \gamma(k) \\ \gamma(k) & \gamma(0) \end{bmatrix}.$$

Since Γ is positive semi-definite it follows that: $\det[A] \geq 0$ so that

$$\gamma(0)^2 - \gamma(k)^2 \geq 0 \implies \gamma(k)^2 \leq \gamma(0)^2$$

or

$$|\gamma(k)| \leq \gamma(0)$$

which proves Theorem 27. ■

Remark 28 *The fact that $\gamma(k)$ is an even function means that if we have calculated $\gamma(k)$ for $k \geq 0$ then we can directly obtain $\gamma(k)$ for $k < 0$ using $\gamma(-k) = \gamma(k)$. For example if $\gamma(3) = 0.7$ then $\gamma(-3) = 0.7$.*

Remark 29 *In general one is on safe ground in simply assuming that Γ in Theorem 25 is positive definite. Suppose that Γ were only positive semi-definite so that:*

$$Var\left[\sum_{i=1}^{k} a_i Y_{t_i}\right] = 0$$

with say $a_1 \neq 0$. Then it would follow that:

$$Y_{t_1} = -\frac{1}{a_1}\sum_{i=2}^{k} a_i Y_{t_i}$$

and so we could make a perfect prediction of Y_{t_1} from the remaining Y_{t_i} 's. Perfect prediction is, aside from accounting identities, something that is pretty rare in economics and so for sensible problems we can simply assert that Γ is positive definite without wasting any energy on the issue.

The AR(1) Model

For the AR(1) model we have already shown that:

$$\gamma(0) = Var[Y_t] = \frac{\sigma^2}{1-\phi^2}$$

and that for $k > 0$:

$$\begin{aligned}\gamma(k) &= \phi^k \gamma(0) \\ &= \phi^k \frac{\sigma^2}{1-\phi^2}.\end{aligned}$$

We can make this formula correct for all k by appealing to Theorem 24 and replacing k with $|k|$ to obtain:

Theorem 30 *For an AR(1) process the autocovariance function is given by:*

$$\gamma(k) = \frac{\phi^{|k|}\sigma^2}{1-\phi^2}.$$

2.5 The Autocorrelation Function

The trouble with covariances is that they generally depend on the units with which Y_t is measured. We can easily get around this problem by working correlations, which are just scaled versions of covariances. We have:

Definition 31 *The correlation between Y_t and Y_{t-k} is:*

$$Corr\left[Y_t, Y_{t-k}\right] = \frac{Cov\left[Y_t, Y_{t-k}\right]}{Var\left[Y_t\right]^{\frac{1}{2}} Var\left[Y_{t-k}\right]^{\frac{1}{2}}}. \tag{2.23}$$

Using stationarity we can simplify this considerably. Since

$$Cov\left[Y_t, Y_{t-k}\right] = \gamma\left(k\right) \tag{2.24}$$

and by stationarity

$$Var\left[Y_t\right]^{\frac{1}{2}} = Var\left[Y_{t-k}\right]^{\frac{1}{2}} = \gamma\left(0\right)^{\frac{1}{2}} \tag{2.25}$$

we have:

$$Corr\left[Y_t, Y_{t-k}\right] = \frac{\gamma(k)}{\gamma(0)}. \tag{2.26}$$

With this in mind we can define the autocorrelation function

$$\rho\left(k\right) = Corr\left[Y_t, Y_{t-k}\right]$$

as follows:

Definition 32 ***Autocorrelation Function:*** *The autocorrelation function $\rho(k)$ is defined as:*

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}.$$

Since the autocorrelation function $\rho(k)$ is just the autocovariance function $\gamma\left(k\right)$ scaled by $\frac{1}{\gamma(0)}$, both $\gamma\left(k\right)$ and $\rho\left(k\right)$ have very similar properties. In particular:

Theorem 33 $\rho(0) = 1$,

Theorem 34 $\rho\left(k\right)\begin{cases} >0 \\ =0 \\ <0 \end{cases}$ *if and only if* $\gamma\left(k\right)\begin{cases} >0 \\ =0 \\ <0 \end{cases}$

Theorem 35 $\rho(-k) = \rho(k)$ *($\rho\left(k\right)$ is an even function)*

Theorem 36 *Let $t_1, t_2, \ldots t_k$ be any k periods then the symmetric $k \times k$ matrix R with i, j^{th} element given by $R_{ij} = \rho\left(t_i - t_j\right)$ or:*

$$R = \begin{bmatrix} 1 & \rho\left(t_1 - t_2\right) & \rho\left(t_1 - t_3\right) & \cdots & \rho\left(t_1 - t_k\right) \\ \rho\left(t_1 - t_2\right) & 1 & \rho\left(t_2 - t_3\right) & \cdots & \rho\left(t_2 - t_k\right) \\ \rho\left(t_1 - t_3\right) & \rho\left(t_2 - t_3\right) & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho\left(t_{k-1} - t_k\right) \\ \rho\left(t_1 - t_k\right) & \rho\left(t_2 - t_k\right) & \cdots & \rho\left(t_{k-1} - t_k\right) & 1 \end{bmatrix}$$

is positive semi-definite.

Theorem 37 *If*

$$Var\left[\sum_{i=1}^{k} a_i Y_{t_i}\right] > 0$$

whenever there is at least one $a_i \neq 0$ then R in Theorem 36 is positive definite.

Theorem 38 $|\rho(k)| \leq 1$.

2.5.1 The Autocorrelation Function of an AR(1) Process

We have :

Theorem 39 *The autocorrelation function of a stationary AR(1) process is:*

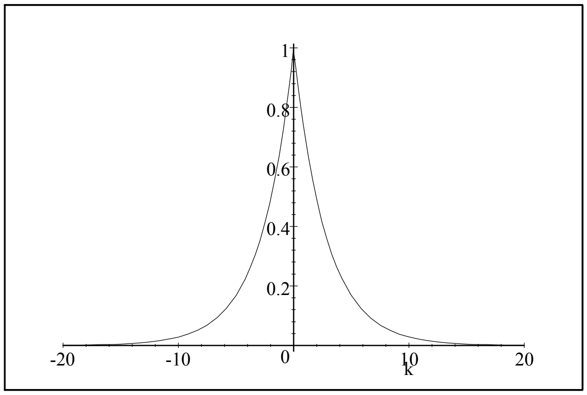
$$\rho(k) = \phi^{|k|}.$$

Proof. From Theorem 30 it follows that

$$\begin{aligned} \rho(k) &= \frac{\gamma\left(k\right)}{\gamma\left(0\right)} \\ &= \frac{\phi^{|k|}\frac{\sigma^2}{1-\phi^2}}{\frac{\sigma^2}{1-\phi^2}} \\ &= \phi^{|k|}. \end{aligned}$$

■

We plot $\rho(k)$ an AR(1) with $\phi = 0.7$ below: [2]



$\rho\left(k\right)$ when $\phi = 0.7$

Since $|\phi| < 1$ it follows that the autocorrelation function, like the autocovariance function, has the short-memory property so that $\rho\left(k\right) = O\left(\tau^k\right)$ as given in Section 2.2 with $A = 1$ and $\tau = |\phi|$.

2.6 Forecasting

2.6.1 The Conditional Mean

Consider the problem of forecasting future Y_{t+k} given the entire history of Y_t up until time t; that is with the information set $I_t = \{Y_t, Y_{t-1}, Y_{t-2}, \dots .\}$. We use the conditional mean $E_t\left[Y_{t+k}\right]$ to forecast Y_{t+k} so that:

Definition 40 *The forecast of Y_{t+k} given the history of Y_t up to time t is the conditional mean $E_t[Y_{t+k}]$ defined as:*

$$\begin{aligned} E_t[Y_{t+k}] &= E\left[Y_{t+k}|I_t\right] \\ &= E\left[Y_{t+k}|Y_t, Y_{t-1}, Y_{t-2}, \dots\right]. \end{aligned}$$

Forecasting an AR(1)

To forecast an AR(1) process we have:

Theorem 41 *For an AR(1) process the optimal forecast k periods in the future is:*

$$E_t[Y_{t+k}] = \phi^k Y_t. \tag{2.27}$$

[2] Note that $\rho\left(k\right)$ is strictly speaking only defined for integer values of k. The plot here essentially connects the values of $\rho\left(k\right)$ between $k = 0, 1, 2, ...$ to produce a continuous plot.

Proof. For an AR(1) process we have shifting (2.6) k periods in the future that:

$$Y_{t+k} = \phi Y_{t+k-1} + a_{t+k}.$$

Applying E_t to both sides we obtain:

$$E_t[Y_{t+k}] = \phi E_t[Y_{t+k-1}] + \underbrace{E_t[a_{t+k}]}_{=0} \tag{2.28}$$

where: $E_t[a_{t+k}] = 0$ since a_{t+k} is *i.i.d.* and hence independent of the information at time t. Hence:

$$E_t[Y_{t+k}] = \phi \underbrace{E_t[Y_{t+k-1}]}_{=\phi E_t[Y_{t+k-2}]}.$$

Continuing this process of substitution we have:

$$E_t[Y_{t+k}] = \phi^k E_t[Y_t]. \tag{2.29}$$

Since Y_t is observed at time t and is thus in the information set, it follows that:

$$E_t[Y_t] = Y_t. \tag{2.30}$$

■

Remark 42 *Note that the forecast function* $E_t[Y_{t+k}] = O\left(\tau^k\right)$ *where* $\tau = |\phi|$ *and hence the forecast function has the short-memory property given in Section 2.2. This means that as we look farther in the future* Y_{t+k} *rapidly forgets the information set at time* t *and so* $E_t[Y_{t+k}]$ *converges rapidly to the unconditional mean* $E[Y_{t+k}] = 0$.

2.6.2 The Conditional Variance

In addition to our forecast we often need some idea of the accuracy of the forecast. This can be determined from the conditional variance defined as follows:

Definition 43 $Var_t[Y_{t+k}]$ *is the conditional variance of* Y_{t+k} *given the information set at time* t *defined as:*

$$Var_t[Y_{t+k}] = Var\left[Y_{t+k}|Y_t, Y_{t-1}, Y_{t-2}, \dots\right].$$

This can be used to construct confidence intervals for our forecasts using the following result:

Theorem 44 *If the stationary stochastic process* Y_t *is Gaussian (normally distributed) a 95% confidence interval for* Y_{t+k} *is:*

$$E_t\left[Y_{t+k}\right] \pm 1.96\sqrt{Var_t[Y_{t+k}]}.$$

The Conditional Variance of an AR(1)

To calculate $Var_t[Y_{t+k}]$ for an AR(1) we have:

Theorem 45 *The value of an AR(1) process k periods in the future: Y_{t+k} can be decomposed into unknown future shocks and the optimal forecast as:*

$$Y_{t+k} = \underbrace{a_{t+k} + \phi^1 a_{t+k-1} + \phi^2 a_{t+k-2} + \cdots + \phi^{k-1} a_{t+1}}_{\text{unknown future shocks}} + \phi^k Y_t.$$

Proof. Note that from (2.6) we can write:

$$\begin{aligned} Y_{t+k} &= \phi Y_{t+k-1} + a_{t+k} \\ &= \phi\left(\phi Y_{t+k-2} + a_{t+k-1}\right) + a_{t+k} \\ &= a_{t+k} + \phi a_{t+k-1} + \phi^2 Y_{t+k-2.} \end{aligned} \tag{2.31}$$

Continuing this we have:

$$Y_{t+k} = a_{t+k} + \phi a_{t+k-1} + \phi^2 a_{t+k-2} + \cdots + \phi^{k-1} a_{t+1} + \phi^k Y_t.$$

■

From this it follows that

Theorem 46 *For an AR(1) process:*

$$\begin{aligned} Var_t[Y_{t+k}] &= \sigma^2(1 + \phi^2 + \phi^4 + \cdots + \phi^{2(k-1)}) \\ &= \frac{\sigma^2(1-\phi^{2k})}{1-\phi^2}. \end{aligned}$$

Proof. The second equality follows from the geometric series:

$$1 + \lambda + \lambda^2 + \cdots + \lambda^{n-1} = \frac{1-\lambda^n}{1-\lambda}$$

when $n = k$ and $\lambda = \phi^2$ and the fact that $Var_t\left[\phi^k Y_t\right] = 0$ since Y_t is in the information set. ■

Remark 47 *Note that*

$$\begin{aligned} Var_t[Y_{t+k}] &= \frac{\sigma^2(1-\phi^{2k})}{1-\phi^2} \\ &= \gamma(0) - \frac{\phi^{2k}\sigma^2}{1-\phi^2} \end{aligned}$$

and so as $k \to \infty$ $Var_t[Y_{t+k}] \to \gamma(0)$.

Since from Y_t is Gaussian we can construct confidence intervals for our forecasts using Theorem 44 as:

Theorem 48 *A 95% confidence interval for the optimal forecast for an AR(1) process is given by:*

$$\phi^k Y_t \pm 1.96 \frac{\sigma\sqrt{1-\phi^{2k}}}{\sqrt{1-\phi^2}}.$$

An Example

Suppose that:

$$Y_t = 0.7Y_{t-1} + a_t \ , \ \ a_t \sim N\left[0, \sigma^2\right] \quad \sigma = 0.03. \tag{2.32}$$

The unconditional distribution is then given by

$$Y_t \sim N\left[0, \frac{(0.03)^2}{1-0.7^2}\right] \tag{2.33}$$

or:

$$Y_t \sim N\left[0, (0.042)^2\right] \tag{2.34}$$

so that 95% of the values of Y_t fall in the band

$$-0.0823 \leq Y_t \leq 0.0823. \tag{2.35}$$

Suppose now that you observe today that $Y_t = 0.02$. Then we have

$$\begin{aligned} E_t[Y_{t+k}] &= (0.7)^k(0.02) \\ Var_t[Y_{t+k}] &= \frac{(0.03)^2(1-(0.7)^{2k})}{(1-(0.7)^2)} \end{aligned} \tag{2.36}$$

so that:

AR(1) Forecasts and Confidence Intervals

k	Forecast $E_t[Y_{t+k}]$	$\sqrt{Var_t[Y_{t+k}]}$	Forecast Confidence Interval
0	0.02	0	0.02 ± 0
1	0.014	0.03	0.014 ± 0.059
2	0.0098	0.037	0.0098 ± 0.072
3	0.0069	0.039	0.0069 ± 0.077
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	0	0.042	0.0 ± 0.082

2.7 The Backward Shift Operator

A very useful device is:

Definition 49 ***The Backward Shift Operator*** *The backward shift operator is defined by:*

$$B^k Y_t \equiv Y_{t-k}.$$

Remark 50 *Note that using this definition:* $B^0 Y_t = Y_t$ *and* $B^{-1} Y_t = Y_{t+1}$ *and in general* $B^{-k} Y_t = Y_{t+k}$.

Example 51 *The AR(1)*

$$Y_t = 0.5 Y_{t-1} + a_t$$

can be re-written as

$$Y_t = 0.5 B Y_t + a_t$$

or as:

$$(1 - 0.5B)\, Y_t = a_t.$$

From the definition $(1 - 0.5B)$ has no meaning when it is separated from Y_t. We will see however that it is useful to think of $(1 - 0.5B)$ as a mathematical object on its own. For example we will see that the stationarity of this AR(1) depends on the root of $(1 - 0.5B) = 0$ or $B = 2$.

2.8 The Wold Representation

2.8.1 Introduction

The Wold representation is a very important and very general result for stationary time series:

Theorem 52 ***Wold Representation:*** *Every stationary time series* Y_t *with* $E[Y_t] = 0$ *and* $Var\,[Y_t] < \infty$ *has an infinite moving average representation:*

$$Y_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \psi_3 a_{t-3} + \cdots \tag{2.37}$$

where a_t *is an uncorrelated series with* $E\,[a_t] = 0$ *and* $Var\,[a_t] = \sigma^2$.

Using the backward shift notation we can write this result in another suggestive form:

Theorem 53 *The Wold representation can be written as*

$$Y_t = \psi(B)\, a_t$$

where $\psi(B)$ is an infinite order polynomial:

$$\begin{aligned}\psi(B) &= 1 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 + \cdots \\ &= \sum_{j=0}^{\infty} \psi_j B^j\end{aligned}$$

where $\psi_o = 1$.

The coefficients ψ_j are sometimes referred to as the *infinite moving average weights* . A moving average process is a weighted sum of past shocks that we will deal with later. For example a first order moving average process or MA(1) would be:

$$Y_t = a_t + \theta a_{t-1} \tag{2.38}$$

so that $\psi_1 = \theta$.

The infinite moving average weights often have interesting economic interpretations when the a_t *'s* are interpreted as say monetary or technological shocks. In this case ψ_k determines the impact of a shock k periods in the past on the Y_t or:

$$\frac{\partial Y_t}{\partial a_{t-k}} = \psi_k. \tag{2.39}$$

2.8.2 Wold Representation for an AR(1)

It is easy to derive the Wold representation for the AR(1) model. From (2.6) we have:

$$Y_t = \phi \underbrace{Y_{t-1}}_{=\phi Y_{t-2}+a_{t-1}} + a_t = a_t + \phi a_{t-1} + \phi^2 \underbrace{Y_{t-2}}_{=\phi Y_{t-3}+a_{t-2}}$$

and continuing this process:

$$Y_t = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \phi^3 a_{t-3} + \cdots$$

so that:

Theorem 54 *For an AR(1) process:*

$$\psi_k = \phi^k.$$

Since Y_t is stationary it follows that $|\phi| < 1$ and hence $\psi_k \to 0$ exponentially or $\psi_k = O\left(\tau^k\right)$ where $\tau = |\phi|$. Thus like $\rho(k)$ and $E_t[Y_{t+k}] : \psi_k$ has the short-memory property given in Section 2.2. This reflects the fact that as $k \to \infty$, Y_t rapidly forgets the effect of past shocks a_{t-k}.

2.8.3 Implications of the Wold Representation

The following results are direct implications of the Wold representation:

Theorem 55 *From the Wold representation in Theorem* 52 *the autocovariance and autocorrelation functions for Y_t are:*

$$\begin{aligned} \gamma(0) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 \qquad (2.40) \\ \gamma(k) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k} \\ \rho(k) &= \frac{\sum_{j=0}^{\infty} \psi_j \psi_{j+k}}{\sum_{j=0}^{\infty} \psi_j^2}. \end{aligned}$$

We can also use the Wold representation to calculate the covariance of Y_t with past shocks a_{t-k}. This will be useful later on for example to derive the variance of an AR(p). In particular we have:

Theorem 56 *From the Wold representation in Theorem* 52 *the covariance between Y_t and a_{t-k} is given by:*

$$E\left[Y_t a_{t-k}\right] = \psi_k \sigma^2.$$

2.8.4 Stationarity and the Wold Representation

A stationary time series must forget the effect of shocks a_{t-k} in the distant past. In economics we might say that for a stationary time series shocks are *transitory.* Mathematically this becomes:

Theorem 57 *The infinite moving average weights of a stationary process must converge to zero or:*

$$\lim_{k \to \infty} \psi_k = 0.$$

Proof. If Y_t is stationary then by Definition 7

$$\gamma(0) = Var[Y_t] < \infty. \qquad (2.41)$$

Now from (2.40) a finite variance implies that:

$$\gamma(0) = Var[Y_t] = \sigma^2(1 + \psi_1^2 + \psi_2^2 + \cdots) < \infty \qquad (2.42)$$

which implies that the $\psi_k{}'s$ must converge to zero.[3] ■

[3] Stationarity also requires that $\gamma(k)$ be finite. However, there are no additional implications to derive from finite covariances since by Theorem 27 we have:

$$|\gamma(k)| \leq \gamma(0) < \infty \qquad (2.43)$$

and so once we establish a finite variance the covariances are automatically finite as well.

Although $\psi_k \to 0$ is necessary for stationarity, it is not sufficient. For example if ψ_k has the long-memory property:

$$\psi_k = \frac{1}{k^\tau} \tag{2.44}$$

it is clear that $\psi_k \to 0$ as long as $\tau > 0$ and that

$$\begin{aligned} \gamma(0) &= \sigma^2\left(1+\psi_1^2+\psi_2^2+\cdots\right) && (2.45) \\ &= \sigma^2\left(1+\frac{1}{1^{2\tau}}+\frac{1}{2^{2\tau}}+\frac{1}{3^{2\tau}}+\cdots\right). && (2.46) \end{aligned}$$

It is a mathematical fact however that[4]

$$\frac{1}{1^z}+\frac{1}{2^z}+\frac{1}{3^z}+\cdots=\infty$$

for $z \leq 1$ so that if:

$$0 < \tau \leq \frac{1}{2}$$

then:

$$\gamma(0) = \infty.$$

Hence such a series would not have a finite variance and hence would not be stationary.

The problem here is that with a long-memory ψ_k does not converge fast enough to zero to insure that the variance is finite. Thus it is not sufficient to show that $\psi_k \to 0$ as $k \to \infty$ in order to show stationarity.

If however $\psi_k = O\left(\tau^k\right)$ and so has the short-memory property (see Section 2.2) then we can be sure that $\gamma(0) < \infty$ and that the series is stationary. In particular we have:

Theorem 58 *If ψ_k has the short-memory property then $\gamma(0) < \infty$ and the process is stationary.*

Proof. This follows from the formula for the geometric series since if $|\psi_k| \leq A\tau^k$ then:

$$\begin{aligned} \gamma(0) &= \sigma^2\sum_{k=0}^{\infty}\psi_k^2 \\ &\leq \sigma^2 A^2\sum_{k=0}^{\infty}\tau^{2k} \\ &= \frac{\sigma^2 A^2}{1-\tau^2} < \infty \end{aligned}$$

since $1-\tau^2 > 0$ if $|\tau| < 1$. ■

[4] For $z = 1$ it can be shown that:

$$1+\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}-\ln(n)$$

converges to $\gamma \approx 0.57$, called Euler's constant.

2.8.5 Forecasting

We can also use the Wold representation to obtain important results for forecasting. Applying Theorem 52 to Y_{t+k} we obtain:

Theorem 59 *The future value of a stationary series Y_{t+k} can be decomposed into unknown future shocks and current and past shocks as:*

$$Y_{t+k} = \underbrace{a_{t+k} + \psi_1 a_{t+k-1} + \psi_2 a_{t+k-2} + \cdots + \psi_{k-1} a_{t+1}}_{unknown\ future\ shocks} + \underbrace{\psi_k a_t + \psi_{k+1} a_{t-1} + \cdots}_{known\ past\ \&\ present\ shocks}$$

By applying $E_t\left[\ \right]$ to both sides of the result in Theorem 59 we can derive the optimal forecast $E_t\left[Y_{t+k}\right]$ as a function of past and present shocks as:

Theorem 60 *$E_t[Y_{t+k}]$ can be expressed as:*

$$E_t[Y_{t+k}] = \sum_{j=0}^{\infty} \psi_{k+j} a_{t-j}.$$

This involves calculating an infinite sum. For the models we will consider there are usually better ways of calculating $E_t[Y_{t+k}]$ than this result.

However by applying $Var_t\left[\right]$ to both sides of Theorem 59 we obtain a very useful formula for determining $Var_t\left[Y_{t+k}\right]$ as:

Theorem 61 *Using the Wold representation $Var_t\left[Y_{t+k}\right]$ can be expressed as:*

$$Var_t[Y_{t+k}] = \sigma^2(1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{k-1}^2).$$

Note that this involves a finite sum and hence is much more useful in practice. Theorem 61 is in fact the basis for all calculations of $Var_t\left[Y_{t+k}\right]$ that we will consider.

2.8.6 Linear and Nonlinear Time Series

The Wold representation only guarantees that the $a_t\ 's$ are *uncorrelated* across time; it does not guarantee that the $a_t\ 's$ are *independent* across time. Recall that independence is a stronger condition than zero covariance; that is, independence implies zero covariance but zero covariance does not imply independence.

Definition 62 *If the $a_t\ 's$ in the Wold representation for Y_t are independent across time we say that Y_t is a linear time series, otherwise we say that Y_t is a nonlinear time series*

For an AR(1) we have a_t as an *i.i.d.* process an hence an AR(1) is a linear time series. In general any series that is Gaussian will be linear since zero correlation implies independence for the normal distribution.

An important example (especially in finance) of a nonlinear time series is an $ARCH\,(1)$ process where:

$$Y_t = Z_t \left(\alpha_o + \alpha_1 Y_{t-1}^2\right)^{1/2} \tag{2.47}$$

and where Z_t is an *i.i.d.* standard normal process. $ARCH$ models are used to model the bursts of volatility that one often observes in financial markets.

As long as $0 < \alpha_1 < 1$, Y_t will be stationary with a Wold representation where $\psi_k = 0$ for $k > 0$ or:

$$Y_t = a_t + 0a_{t-1} + 0a_{t-2} + \cdots \tag{2.48}$$

and where:

$$a_t = Z_t \left(\alpha_o + \alpha_1 Y_{t-1}^2\right)^{1/2} \tag{2.49}$$

is uncorrelated across time. Although the $a_t's$ are uncorrelated across time they are not independent since for example a_t^2 is correlated with a_{t-1}^2.

2.9 The Yule-Walker Equations

2.9.1 Derivation of the Yule-Walker Equations

The optimal forecast of Y_{t+1} given knowledge of the entire past history of the time series is $E_t\left[Y_{t+1}\right]$. This forecast uses the entire past history of Y_t into the infinite past; that is:

$$I_t = \{Y_t, Y_{t-1}, Y_{t-2}, \dots\}. \tag{2.50}$$

Suppose instead we were only to use the most recent k values in the information set or:

$$I_t^k = \{Y_t, Y_{t-1}, Y_{t-2}, \dots Y_{t-k+1}\} \tag{2.51}$$

and we wish to calculate:

$$E\left[Y_{t+1}|I_t^k\right]. \tag{2.52}$$

What would be the optimal forecast given this information set? This turns out to be an interesting problem because it leads to the partial autocorrelation function used in Box-Jenkins identification.

It turns out that given normality the optimal forecast will be a linear function of the information set so that:

$$E\left[Y_{t+1}|I_t^k\right] = \sum_{j=0}^{k-1} \phi_{jk} Y_{t-j} = X_t^T \phi_k \tag{2.53}$$

where the $k \times 1$ vectors X_t and ϕ_k are given by:

$$X_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-k-1} \end{bmatrix}, \quad \phi_k = \begin{bmatrix} \phi_{1k} \\ \phi_{2k} \\ \vdots \\ \phi_{kk} \end{bmatrix}. \tag{2.54}$$

It can be shown that ϕ_k will minimize the forecast variance:

$$\begin{aligned} S(\phi_k) &= Var\left[\left(Y_{t+1} - E\left[Y_{t+1}|I_t^k\right]\right)\right] \qquad (2.55) \\ &= E\left[\left(Y_{t+1} - E\left[Y_{t+1}|I_t^k\right]\right)^2\right] \\ &= E\left[\left(Y_{t+1} - X_t^T\phi_k\right)^2\right]. \end{aligned}$$

Differentiating $S(\phi_k)$ with respect to ϕ_k and setting this equal to zero yields the first-order conditions:

$$E\left[X_t\left(Y_{t+1} - X_t^T\phi_k\right)\right] = 0 \tag{2.56}$$

which states that the information set X_t must be uncorrelated with the forecast error: $Y_t - X_t^T\phi$. This in turn implies that:

$$\underbrace{E\left[X_tX_t^T\right]}_{\Gamma_k}\phi_k = \underbrace{E\left[X_tY_t\right]}_{g_k} \tag{2.57}$$

where the $k \times k$ symmetric matrix Γ_k is given by:

$$\Gamma_k \equiv E\left[X_tX_t^T\right] = \gamma(0)R_k \tag{2.58}$$

and the symmetric $k \times k$ matrix R_k is:

$$R_k = \begin{bmatrix} 1 & \rho(1) & \rho(2) & \cdots & \rho(k-1) \\ \rho(1) & 1 & \rho(1) & \cdots & \rho(k-2) \\ \rho(2) & \rho(1) & 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \rho(1) \\ \rho(k-1) & \rho(k-2) & \cdots & \rho(1) & 1 \end{bmatrix}. \tag{2.59}$$

Similarly we can show that $k \times 1$ vector g_k is given by:

$$g_k \equiv E\left[X_tY_{t+1}\right] = \gamma(0)r_k \tag{2.60}$$

where:

$$r_k = \begin{bmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(k) \end{bmatrix}. \tag{2.61}$$

We thus have:

$$\gamma(0)\, R_k \phi_k = \gamma(0)\, r_k \tag{2.62}$$

so that cancelling the scalar $\gamma(0)$ from both sides we have:

$$R_k \phi_k = r_k. \tag{2.63}$$

which are the Yule-Walker equations. If R_k is nonsingular we can solve these equations for ϕ_k, the optimal forecast weights. We have:

Theorem 63 *The matrix R_k is positive definite and so R_k^{-1} exists.*

Proof. This follows from Theorem 37 by setting $t_i = t - i$ for $i = 0, 1, 2, \dots k - 1$. ■

It therefore follows that:

Theorem 64 ***Yule-Walker:*** *The vector of optimal forecast weights ϕ_k in $E\left[Y_{t+1}|I_t^k\right] = X_t^T \phi_k$ is given by:*

$$\phi_k = R_k^{-1} r_k$$

Remark 65 *Note that to actually calculate ϕ_k we need R_k which depends on $\rho(1), \rho(2), \dots \rho(k-1)$ and r_k which depends on $\rho(1), \rho(2), \dots \rho(k)$. Thus knowledge of the first k autocorrelations is sufficient to calculation ϕ_k.*

Remark 66 *It can be shown that:*

$$\begin{aligned} Var\left[Y_{t+1}|I_t^k\right] &= \gamma(0) - \gamma(0)\, \phi_k^T R_k \phi_k \\ &= \gamma(0)\left(1 - r_k^T R_k^{-1} r_k\right) \\ &= Var\left[Y_{t+1}\right]\left(1 - r_k^T R_k^{-1} r_k\right) \end{aligned}$$

Thus including the information set I_t^k reduces the unconditional forecast variance $Var\left[Y_{t+1}\right] = \gamma(0)$ by a factor of $\left(1 - r_k^T R_k^{-1} r_k\right)$.

2.9.2 Examples

Forecasts when $k = 1$

For $k = 1$ we have $R_1 = 1$ and $g_k = \rho(1)$ so that from Theorem 64 we have:

$$\phi_{11} = \rho(1). \tag{2.64}$$

Thus the optimal forecast of Y_{t+1} based only on Y_t is:

$$E\left[Y_{t+1}|Y_t\right] = \phi_{11} Y_t = \rho(1) Y_t. \tag{2.65}$$

If for example we have an AR(1): $Y_t = \phi Y_{t-1} + a_t$ then $\rho(1) = \phi$ and

$$E\left[Y_{t+1}|Y_t\right] = \phi Y_t \tag{2.66}$$

which is equal to the *optimal* forecast since:

$$E_t\left[Y_{t+1}\right] = \phi Y_t; \tag{2.67}$$

that is for an AR(1) process a forecast of Y_{t+1} based only on Y_t is just as good as a forecast based on the entire past history of Y_t at time t :

$$Y_t, Y_{t-1}, Y_{t-2}, \ldots .$$

This is a special property of AR processes called the Markov property.

Forecasts when $k = 2$

For $k = 2$ we have from (2.63):

$$\begin{bmatrix} 1 & \rho(1) \\ \rho(1) & 1 \end{bmatrix} \begin{bmatrix} \phi_{12} \\ \phi_{22} \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \rho(2) \end{bmatrix} \tag{2.68}$$

so that using Cramer's rule:

$$\begin{aligned} \phi_{12} &= \frac{\rho(1)(1-\rho(2))}{1-\rho(1)^2} \\ \phi_{22} &= \frac{\rho(2)-\rho(1)^2}{1-\rho(1)^2}. \end{aligned} \tag{2.69}$$

For example given an AR(1) :

$$Y_t = \phi Y_{t-1} + a_t \tag{2.70}$$

we have from Theorem 39 that: $\rho(1) = \phi$ and $\rho(2) = \phi^2$ so that:

$$\begin{aligned} \phi_{12} &= \frac{\phi\left(1-\phi^2\right)}{1-\phi^2} = \phi \\ \phi_{22} &= \frac{\phi^2-\phi^2}{1-\phi^2} = 0. \end{aligned}$$

Thus:

$$\begin{aligned} E\left[Y_{t+1}|Y_t, Y_{t-1}\right] &= \phi Y_t + 0 \times Y_{t-1} \\ &= \phi Y_t. \end{aligned} \tag{2.71}$$

This is again the Markov property of an AR(1) process: the information set at time t is completely summarized by Y_t so that the optimal forecast does not depend on $Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots$.

A Numerical Example

Suppose we wish to use the Yule-Walker equations to derive a forecast rule for $k = 2$ or using

$$E\left[Y_{t+1}|Y_t, Y_{t-1}\right] = \phi_{12}Y_t + \phi_{22}Y_{t-1}. \tag{2.72}$$

Suppose we know that $\rho(1) = 0.6$ and $\rho(2) = 0.5$. Then from the Yule-Walker equations:

$$\begin{bmatrix} 1 & \rho(1) \\ \rho(1) & 1 \end{bmatrix} \begin{bmatrix} \phi_{12} \\ \phi_{22} \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \rho(2) \end{bmatrix} \tag{2.73}$$

or:

$$\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix} \begin{bmatrix} \phi_{12} \\ \phi_{22} \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.5 \end{bmatrix}. \tag{2.74}$$

Solving we find that:

$$\phi_{12} = 0.47 \text{ and } \phi_{22} = 0.22. \tag{2.75}$$

We then have:

$$E\left[Y_{t+1}|Y_t, Y_{t-1}\right] = 0.47Y_t + 0.22Y_{t-1}. \tag{2.76}$$

If $Y_t = 0.03$ and $Y_{t-1} = 0.02$ then our forecast of Y_{t+1} would be :

$$E\left[Y_{t+1}|Y_t, Y_{t-1}\right] = (0.47)(0.03) + (0.22)(0.02) = 0.0185. \tag{2.77}$$

2.9.3 Recursive Calculation of ϕ_k

Once ϕ_k has been calculated this can be used to calculate ϕ_{k+1} recursively using

Proposition 67 ***Durbin's Formula:*** *Given $\phi_k = \left[\phi_{kj}\right]$ then the elements of $\phi_{k+1} = \left[\phi_{k+1,j}\right]$ can be calculated as follows:*

$$\begin{aligned} \phi_{k+1,j} &= \phi_{k,j} - \phi_{k+1,k+1}\phi_{k,k-j+1}, \text{ for } j = 1, 2, \ldots k \\ \phi_{k+1,k+1} &= \frac{\rho(k+1) - \sum_{j=1}^{k} \rho(k-j+1)\,\phi_{k,j}}{1 - \sum_{j=1}^{k} \rho(j)\,\phi_{k,j}}. \end{aligned}$$

2.9.4 The Partial Autocorrelation Function

An important by-product from the Yule-Walker equations is the partial autocorrelation function.

Definition 68 ***Partial Autocorrelation Function:*** *The partial autocorrelation function for a stationary process Y_t is defined as ϕ_{kk} where ϕ_{kk} is the k^{th} element of ϕ_k from Theorem* 64

$$\phi_k = R_k^{-1} r_k.$$

From Cramer's rule and the Yule-Walker equations we have:

$$\begin{aligned}
\phi_{11} &= \rho\left(1\right), \qquad (2.78)\\
\phi_{22} &= \frac{\begin{vmatrix} 1 & \rho\left(1\right) \\ \rho\left(1\right) & \rho\left(2\right) \end{vmatrix}}{\begin{vmatrix} 1 & \rho\left(1\right) \\ \rho\left(1\right) & \rho\left(1\right) \end{vmatrix}} = \frac{\rho\left(2\right) - \rho\left(1\right)^2}{1 - \rho\left(1\right)^2},\\
\phi_{33} &= \frac{\begin{vmatrix} 1 & \rho\left(1\right) & \rho\left(1\right) \\ \rho\left(1\right) & 1 & \rho\left(2\right) \\ \rho\left(2\right) & \rho\left(1\right) & \rho\left(3\right) \end{vmatrix}}{\begin{vmatrix} 1 & \rho\left(1\right) & \rho\left(2\right) \\ \rho\left(1\right) & 1 & \rho\left(1\right) \\ \rho\left(2\right) & \rho\left(1\right) & 1 \end{vmatrix}} = \frac{\rho\left(3\right)\left(1 - \rho\left(1\right)^2\right) - 2\rho\left(2\right)\rho\left(1\right) + \rho\left(1\right)^3 + \rho\left(2\right)^2\rho\left(1\right)}{1 - 2\rho\left(1\right)^2 + 2\rho\left(2\right)\rho\left(1\right)^2 - \rho\left(2\right)^2}.\\
&\text{etc..}
\end{aligned}$$

Note that ϕ_{kk} is a function of $\rho\left(1\right), \rho\left(2\right), \ldots \rho\left(k\right)$.

Example 69 *Suppose that:*

$$\rho\left(1\right) = 0.8, \rho\left(2\right) = 0.5, \rho\left(3\right) = 0.3$$

then

$$\begin{aligned}
\phi_{11} &= 0.8,\\
\phi_{22} &= \frac{\begin{vmatrix} 1 & 0.8 \\ 0.8 & 0.5 \end{vmatrix}}{\begin{vmatrix} 1 & 0.8 \\ 0.8 & 1 \end{vmatrix}} = -0.39,\\
\phi_{33} &= \frac{\begin{vmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.5 \\ 0.5 & 0.8 & 0.3 \end{vmatrix}}{\begin{vmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0.8 \\ 0.5 & 0.8 & 1 \end{vmatrix}} = 0.18.
\end{aligned}$$

ϕ_{kk} is called the partial autocorrelation function because it can be shown to be equal to the correlation between Y_{t+1} and Y_{t-k+1} when the predictive effects of the intermediate values $I_t^k = \{Y_t, Y_{t-1}, Y_{t-2}, \ldots Y_{t-k+1}\}$ are removed; that is:

Theorem 70

$$\phi_{kk} = Corr\left[Y_{t+1}, Y_{t+1-k} | I_t^k\right]. \qquad (2.79)$$

Proof. Recall that for any two scalar random variables: X_1 and X_2 that are jointly normally distributed that:

$$E\left[X_1 | X_2\right] = E\left[X_1\right] + \frac{Cov\left[X_1, X_2\right]}{Var\left[X_2\right]}\left(X_2 - E\left[X_2\right]\right).$$

It follows then that:

$$E\left[Y_{t+1}|Y_{t+1-k,}I_t^k\right]=E\left[Y_{t+1}|I_t^k\right]+\frac{Cov\left[Y_{t+1},Y_{t+1-k}|I_t^k\right]}{Var\left[Y_{t+1-k}|I_t^k\right]}\left(Y_{t+1-k}-E\left[Y_{t+1-k}|I_t^k\right]\right).$$

Since the coefficient on Y_{t+1-k} is ϕ_{kk} it follows that:

$$\phi_{kk}=\frac{Cov\left[Y_{t+1},Y_{t+1-k}|I_t^k\right]}{Var\left[Y_{t+1-k}|I_t^k\right]}.$$

By symmetry (or as an exercise):

$$Var\left[Y_{t+1-k}|I_t^k\right]=Var\left[Y_{t+1}|I_t^k\right]$$

and so:

$$\phi_{kk}=\frac{Cov\left[Y_{t+1},Y_{t+1-k}|I_t^k\right]}{Var\left[Y_{t+1}|I_t^k\right]^{\frac{1}{2}}Var\left[Y_{t+1-k}|I_t^k\right]^{\frac{1}{2}}}.$$

■

Thus $\phi_{11}=\rho(1)$ because there are no intermediate values between Y_t and Y_{t-1}. In general though $\phi_{22}\neq\rho(2)$ since when we remove the predictive effects of Y_{t-1} when determining the correlation between Y_t and Y_{t-2} we get a different correlation.

Since ϕ_{kk} is a correlation it is bounded between -1 and 1 and so we have:

Theorem 71

$$-1\leq\phi_{kk}\leq1. \tag{2.80}$$

Chapter 3

AR(p) Processes

3.1 Introduction

We can generalize a first-order autoregressive process to a p^{th} order as follows:

Definition 72 ***AR(p) Processes:*** *We say that Y_t follows an AR(p) process or Y_t ∼AR(p) if :*

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + a_t$$

where $a_t \sim i.i.n(0, \sigma^2)$.

An important question will be the conditions under which an AR(p) process is stationary. These are somewhat counter-intuitive. For example:

$$Y_t = 0.7Y_{t-1} + 0.35Y_{t-2} + a_t \tag{3.1}$$

turns out to be nonstationary despite the fact that $|\phi_1| < 1$ and $|\phi_2| < 1$ while

$$Y_t = 1.4Y_{t-1} - 0.7Y_{t-2} + a_t \tag{3.2}$$

turns out to be stationary despite the fact that $\phi_1 > 1$.

To obtain the conditions for the stationarity of an AR(p) process we will require the backward shift operator B.

We can use the backward shift operator B to rewrite the AR(p) process much more compactly. Since:

$$Y_t = \phi_1 \underbrace{Y_{t-1}}_{=BY_t} + \phi_2 \underbrace{Y_{t-2}}_{=B^2Y_t} + \cdots + \phi_p \underbrace{Y_{t-p}}_{=B^pY_t} + a_t \tag{3.3}$$

it follows that:

$$Y_t = \phi_1 BY_t + \phi_2 B^2 Y_t + \cdots + \phi_p B^p Y_t + a_t \tag{3.4}$$

or placing all the terms involving Y_t on the left-hand side and factoring out Y_t :

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)Y_t=a_t. \tag{3.5}$$

We therefore have:

Theorem 73 *An AR(p) can be written as:*

$$\phi(B)Y_t=a_t \tag{3.6}$$

where $\phi(B)$ is a p^{th} order polynomial in B given by:

$$\phi(B)=1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p. \tag{3.7}$$

Example 74 *The $AR(2)$ process*

$$Y_t=0.6Y_{t-1}+0.2Y_{t-2}+a_t \tag{3.8}$$

can be rewritten with

$$\phi(B)=1-0.6B-0.2B^2 \tag{3.9}$$

as:

$$\left(1-0.6B-0.2B^2\right)Y_t=a_t.$$

3.2 Some Derivations Using B

3.2.1 Wold Representation of an AR(1)

The backward shift operator might now be simply thought of as an empty notational convention, but this notation turns out to be a very fruitful one. For example consider the following alternative method for deriving the Wold representation for the $AR(1)$ model. Throw $(1-\phi B)$ on the left-hand side onto the right-hand side as:

$$Y_t=\frac{1}{1-\phi B}a_t. \tag{3.10}$$

Since $Y_t=\psi(B)a_t$ we have:

$$\psi(B)=\frac{1}{1-\phi B}=1+\phi B+\phi^2B^2+\phi^3B^3+\cdots \tag{3.11}$$

using the geometric series which states that:

$$\frac{1}{1-x}=1+x+x^2+x^3+\cdots. \tag{3.12}$$

3.2.2 Wold Representation for an AR(2)

Now consider an $AR(2)$ process:

$$Y_t = 1.3Y_{t-1} - 0.4Y_{t-2} + a_t \tag{3.13}$$

which can in turn be rewritten as:

$$\phi(B)Y_t = a_t$$

where:

$$\phi(B) = 1 - 1.3B + 0.4B^2.$$

We might then ask, is it legitimate to throw $\phi\left(B\right)$ onto the right-hand side as we did for the AR(1) model to obtain the Wold representation? That is can we go from (3.13) to:

$$Y_t = \frac{1}{1 - 1.3B + 0.4B^2} a_t.$$

This answer is yes, provided that $\phi\left(B\right)$ has certain properties related to its roots. We can factor $\phi\left(B\right)$ as:

$$\begin{aligned} \phi(B) &= 1 - 1.3B + 0.4B^2 \\ &= \left(1 - 0.5B\right)\left(1 - 0.8B\right). \end{aligned}$$

Note that the roots of $\phi\left(B\right)$ are the inverses of 0.5 and 0.8, the coefficients on B in the factorization. You can verify then that:

$$\frac{1}{1 - 1.3B + 0.4B^2} = \frac{-\frac{5}{3}}{(1 - 0.5B)} + \frac{\frac{8}{3}}{(1 - 0.8B)}$$

so that:

$$\begin{aligned} Y_t &= \frac{1}{1 - 1.3B + 0.4B^2} a_t \\ &= \left(\frac{-\frac{5}{3}}{(1 - 0.5B)} + \frac{\frac{8}{3}}{(1 - 0.8B)}\right) a_t \\ &= \sum_{k=0}^{\infty} \left(-\frac{5}{3}\left(0.5\right)^k + \frac{8}{3}\left(0.8\right)^k\right) a_{t-k} \end{aligned}$$

where the last line follows from the geometric series. We conclude that the Wold representation for Y_t is given by:

$$\psi_k = \left(-\frac{5}{3}\left(0.5\right)^k + \frac{8}{3}\left(0.8\right)^k\right)$$

and that $\psi_k \rightarrow 0$ as $k \rightarrow \infty$.

3.3 The Wold Representation for an AR(p) Process

Suppose that $Y_t \sim AR(p)$. Given that Y_t is stationary it has a Wold representation:

$$Y_t = \psi(B)a_t. \tag{3.14}$$

Since $\phi(B)Y_t = a_t$ we have

$$\phi(B)Y_t = a_t = \phi(B)\psi(B)a_t \tag{3.15}$$

so that:

$$\phi(B)\psi(B) = 1 \tag{3.16}$$

or equivalently:

$$\phi(B)\psi(B) = 1 + 0B + 0B^2 + 0B^3 + \cdots . \tag{3.17}$$

Writing this out explicitly we have:

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(\psi_0 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 + \cdots) = 1 + 0B + 0B^2 + \cdots \tag{3.18}$$

and equating the coefficient on $B^0 = 1$ on both sides we have:

$$\psi_0 = 1 \tag{3.19}$$

which we already knew.

Equating coefficients on B^1 on both sides we have:

$$(\psi_1 - \phi_1\psi_0)\, B^1 = 0B^1$$

or

$$\psi_1 - \phi_1\psi_0 = 0$$

or using the fact that $\psi_0 = 1$ that:

$$\psi_1 = \phi_1.$$

We can do the same thing with the coefficients on B^2 which yields:

$$\left(\psi_2 - \phi_1 \underbrace{\psi_1}_{=\phi_1} - \phi_2 \underbrace{\psi_0}_{=1}\right) \times B^2 = 0 \times B^2 \tag{3.20}$$

or:

$$\psi_2 - \phi_1\psi_1 - \phi_2\psi_0 = 0$$

or:

$$\begin{aligned}\psi_2 &= \phi_1\psi_1 + \phi_2\psi_0 &(3.21)\\ &= \phi_1^2 + \phi_2. &(3.22)\end{aligned}$$

If you keep on doing this you will that for B^k:

$$\left(\psi_k - \phi_1\psi_{k-1} - \phi_2\psi_{k-2} - \cdots - \phi_p\psi_{k-p}\right)B^k = 0 \times B^k$$

so that we obtain the following theorem:[1]

Theorem 75 ***Recursive Calculation of*** ψ_k*: For a stationary AR(p) process* $\phi(B)Y_t = a_t$ *the Wold representation can be recursively calculated as:*

$$\psi_k = \phi_1\psi_{k-1} + \phi_2\psi_{k-2} + \cdots + \phi_p\psi_{k-p}$$

with starting values:

$$\psi_0 = 1,\ \psi_k = 0 \textit{ for } k < 0.$$

An Example

For example consider calculating the Wold representation for the $AR(2)$ model:

$$Y_t = 1.4Y_{t-1} - 0.7Y_{t-2} + a_t. \qquad (3.23)$$

From Theorem 75 we have:

$$\psi_k = 1.4\psi_{k-1} - 0.7\psi_{k-2}, \qquad (3.24)$$

with starting values:

$$\psi_0 = 1,\ \psi_{-1} = 0 \qquad (3.25)$$

which can now be used to calculate ψ_k recursively. Thus:

$$\psi_1 = 1.4\underbrace{\psi_0}_{=1} - 0.7\underbrace{\psi_{-1}}_{=0} = 1.4 \qquad (3.26)$$

[1] Another way of stating this result is that for $k > 0$:

$$\phi(B)\psi_k = 0$$

where the backward shift operator works on the k subscript so that for example:

$$B^2\psi_k = \psi_{k-2}.$$

$$\psi_2 = 1.4 \underbrace{\psi_1}_{=1.4} - 0.7 \underbrace{\psi_0}_{=1} = 1.26$$

$$\psi_3 = 1.4 \underbrace{\psi_2}_{=1.26} - 0.7 \underbrace{\psi_1}_{=1.4} = 0.784$$

$$\begin{aligned}\psi_4 &= 1.4 \underbrace{\psi_3}_{=0.784} - 0.7 \underbrace{\psi_2}_{=1.26} = 0.215, \\ &\text{etc..}\end{aligned}$$

If you continue with this process you will find that $\psi_k \to 0$ very rapidly as k gets large which reflects the fact that this particular AR(2) is stationary.

3.4 Stationarity Conditions for an AR(p)

For an AR(p) we have seen from Theorem 75 how ψ_k can be calculated recursively. We know from Theorem 57 that if Y_t is stationary then ψ_k must converge to zero. Therefore an empirical method of determining stationarity would be to calculate ψ_k for k large using Theorem 75 and see if appears to be converging to zero.

We can however settle the issue of stationary by using the fact that:

$$\begin{aligned}\psi_k &= \sum_{j=1}^{p} \phi_j \psi_{k-j} \qquad (3.27) \\ \psi_0 &= 1, \ \psi_k = 0 \ \text{ for } k < 0.\end{aligned}$$

from Theorem 75 and so ψ_k follows a p^{th} order linear difference equation. We solve linear difference equations in the usual way by conjecturing a solution of the form:

$$\psi_k = Ar^k \qquad (3.28)$$

which leads to:

$$\underbrace{Ar^k}_{\psi_k} = \phi_1 \underbrace{Ar^{k-1}}_{\psi_{k-1}} + \phi_2 \underbrace{Ar^{k-2}}_{\psi_{k-2}} + \cdots + \phi_p \underbrace{Ar^{k-p}}_{\psi_{k-p}}. \qquad (3.29)$$

Cancelling A and r^k from both sides we have:

$$1 = \phi_1 r^{-1} + \phi_2 r^{-2} + \cdots + \phi_p r^{-p} \qquad (3.30)$$

or

$$1 - \phi_1 r^{-1} - \phi_2 r^{-2} - \cdots - \phi_p r^{-p} = 0. \qquad (3.31)$$

Now since

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{3.32}$$

this can be written compactly as

$$\phi(r^{-1}) = 0; \tag{3.33}$$

that is, if we replace B by r^{-1} in $\phi(B)$ we get zero.

It follows that r^{-1} is a root of $\phi(B)$. Since $\phi(B)$ is a p^{th} order polynomial we will have p roots:

$$r_1^{-1}, r_2^{-1}, \ldots r_p^{-1} \tag{3.34}$$

so that

Theorem 76 *For an AR(p) the infinite moving average weights ψ_k can be expressed as:*

$$\psi_k = A_1 r_1^k + A_2 r_2^k + \cdots + A_p r_p^k$$

where: r_i^{-1} is a root of $\phi(B)$; that is: $\phi(r_i^{-1}) = 0$ and A_i can be found from the starting values

$$\psi_0 = 1,\ \psi_k = 0 \ \textit{ for } k < 0.$$

It is slightly awkward that r^{-1} is the root of $\phi(B)$ and not r. This can be remedied by defining a closely related p^{th} order polynomial where we begin with a power of p on the left-hand side instead of the power 0 as we do with $\phi(B)$. More formally:

Definition 77 *Let the p^{th} order polynomial $\tilde{\phi}(r)$ be defined as:*

$$\begin{aligned} \tilde{\phi}(r) &= r^p \phi(r^{-1}) \\ &= r^p - \phi_1 r^{p-1} - \phi_2 r^{p-2} - \phi_3 r^{p-3} - \cdots - \phi_p. \end{aligned}$$

We then have:

Theorem 78 *r_i^{-1} is a root of $\phi(B)$ if and only if r_i is a root of $\tilde{\phi}(r)$.*

Proof. If $\phi\left(r_i^{-1}\right) = 0$ then $\tilde{\phi}(r_i) = r_i^p \phi(r_i^{-1}) = 0$. Similarly if $\tilde{\phi}(r_i) = 0$ then $r_i^p \phi(r_i^{-1}) = 0$ or $\phi(r_i^{-1}) = 0$. ■

Example 79 *Given*

$$\phi(B) = 1 - 0.5B - 0.3B^2 \tag{3.35}$$

which has roots

$$B = -2.8403 \textit{ and } B = 1.1736. \tag{3.36}$$

$\tilde{\phi}(r)$ is then given by:

$$\tilde{\phi}(r) = r^2 - 0.5r - 0.3 \tag{3.37}$$

which has roots:

$$\begin{aligned} r_1 &= -0.35208 = \frac{1}{-2.8403} \text{ and} \\ r_2 &= 0.85208 = \frac{1}{1.1736}. \end{aligned} \tag{3.38}$$

Since for a stationary process we have from Theorem 57 that:

$$\lim_{k \to \infty} \psi_k = 0 \tag{3.39}$$

an AR(p) process will only be stationary if for each r_i^{-k} as $k \to \infty$:

$$r_i^{-k} \to 0. \tag{3.40}$$

This in turn will only occur if:

$$|r_i| < 1 \tag{3.41}$$

for *all* roots.

We can therefore have the essential condition for stationarity:

Theorem 80 *An AR(p) process: $\phi(B)Y_t = a_t$ is stationary if and only if all roots of $\phi(B)$ are greater than 1 in absolute value; that is if for all $i = 1, 2, \ldots p$*

$$\phi\left(r_i^{-1}\right) = 0 \Longrightarrow |r_i| < 1.$$

We can also express the stationarity condition in terms of the $\tilde{\phi}(r)$ as:

Theorem 81 *An AR(p) process is stationary if an only if all roots of $\tilde{\phi}(r)$ given in Definition 77 are less than 1 in absolute value; that is if for all $i = 1, 2, \ldots p$*

$$\tilde{\phi}(r_i) = 0 \Longrightarrow |r_i| < 1.$$

Remark 82 *We have only proven necessity. For practical purposes these two conditions are also* sufficient *for stationarity. To be precise however we should make some qualification regarding the distribution of the starting values of the process. For example if for a stationary AR(1) it were the case that the starting value Y_o had say a t distribution instead of a normal, then $Y_t = \phi Y_{t-1} + a_t$ would not be strictly speaking stationary since it would take some time for Y_t to forget the distribution of Y_o. We would then say that Y_t is* asymptotically *stationary. Theoretical models often assume that Y_t has an infinite past in order to get around this problem.*

Note that given stationarity $r_i^k \to 0$ exponentially. This would suggest that ψ_k for a stationary AR(p) has the *short-memory* property in Section 2.2. We have:

Theorem 83 *For a stationary AR(p) process ψ_k has the short-memory property: $\psi_k = O\left(\tau^k\right)$ or*

$$|\psi_k| \leq A\tau^k$$

where

$$0 \leq \tau = \max_i \left[|r_i|\right] < 1.$$

Proof. From the Cauchy-Schwarz inequality we have:

$$\begin{aligned} |\psi_k| &= \left|\sum_{i=1}^{p} A_i r_i^k\right| \\ &\leq \left(\sum_{i=1}^{p} A_i^2\right)^{1/2} \left(\sum_{i=1}^{p} |r_i|^{2k}\right)^{1/2} \\ &\leq A\tau^k \end{aligned}$$

where

$$0 \leq \tau = \max_i \left[|r_i|\right] < 1$$

is the absolute value of the largest root of $\tilde{\phi}\left(r\right)$ and:

$$A = \left(\sum_{i=1}^{p} A_i^2\right)^{1/2}.$$

■

From this short-memory result it follows that:

Theorem 84 *If for the AR(p) process $\phi\left(B\right)Y_t = a_t$, $\phi\left(B\right)$ has all roots greater than one in absolute value or $|r_i| < 1$ for $i = 1, 2, \ldots p$ then:*

$$|\gamma\left(k\right)| \leq \gamma\left(0\right) < \infty.$$

Example 1

The process

$$Y_t = 0.5Y_{t-1} + 0.2Y_{t-2} + a_t$$

is stationary since

$$\tilde{\phi}(r) = r^2 - 0.5r - 0.2 = 0 \tag{3.42}$$

implies that:

$$r = \frac{0.5 \pm \sqrt{(0.5)^2 - 4(-0.2)}}{2} \tag{3.43}$$

or $r_1 = 0.762$ and $r_2 = -0.262$. Thus since $|r_1| < 1$ and $|r_2| < 1$ we conclude that the process is stationary.

Example 2

The process $Y_t = 0.8Y_{t-1} + 0.3Y_{t-2} + a_t$ is not stationary since

$$\tilde{\phi}(r) = r^2 - 0.8r - 0.3 = 0 \tag{3.44}$$

implies that:

$$r = \frac{0.8 \pm \sqrt{(0.8)^2 - 4(-0.3)}}{2} \tag{3.45}$$

or $r_1 = 1.08$ and $r_2 = -0.278$. Thus since $|r_1| = 1.08 > 1$ we conclude that the process is not stationary.

3.4.1 Necessary Conditions for Stationarity

Sometimes it is not necessary to calculate the roots of a polynomial to know that a process is nonstationary. Here we derive a number of necessary conditions for stationarity; that is if these conditions are violated then we know the process is nonstationary but we cannot conclude from their being satisfied that the process is stationary.

Consider the factorization of the p^{th} order polynomial

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{3.46}$$

as:

$$\phi(B) = (1 - r_1 B)(1 - r_2 B) \cdots (1 - r_p B). \tag{3.47}$$

Multiplying these terms out and equating the coefficient on B^p in the two representations it follows that we have:

$$|\phi_p| = |r_1||r_2| \cdots |r_p|. \tag{3.48}$$

Since the process is stationary we have: $|r_i| < 1$ for $i = 1, 2, \ldots p$ and so we conclude that:

Theorem 85 *A necessary condition for the stationarity of an AR(p) process is that:*

$$|\phi_p| < 1. \tag{3.49}$$

Thus the coefficient on the last lag Y_{t-p} must always be less than 1 in absolute value.

Example 86 *We can say that*

$$Y_t = 0.1Y_{t-1} - 1.1Y_{t-2} + a_t$$

is not stationary since $|\phi_2| = 1.1 > 1$ *without bothering to calculate the roots.*

Another necessary condition for stationarity is that the sum of the ϕ_i *'s* must sum to less than 1 or:

Theorem 87 *A necessary condition for the stationarity of an AR(p) process is that* $\phi(1) > 0$ *or:*

$$\phi_1 + \phi_2 + \cdots + \phi_p < 1. \tag{3.50}$$

Proof. If r_i is real and $|r_i| < 1$ then clearly:

$$1 - r_i > 0.$$

Thus if all roots are real

$$\phi(1) = (\overset{+}{1 - r_1})(\overset{+}{1 - r_2}) \cdots (\overset{+}{1 - r_p}) > 0 \tag{3.51}$$

The same argument applies with complex roots. Thus if there is a complex root, say $r_k = a + bi$, then it will be paired with another root will which is its complex conjugate say $r_l = a - bi$ in which case:

$$\begin{aligned} (1 - r_k)(1 - r_l) &= (1 - (a + bi))(1 - (a - bi)) \\ & \quad 1 - 2a + a^2 + b^2 \\ &= (1 - a)^2 + b^2 > 0 \end{aligned}$$

and so $\phi(1) > 0$. Now since

$$\phi(1) = 1 - \phi_1 - \phi_2 - \cdots - \phi_p > 0 \tag{3.52}$$

the result follows. ■

Example 88 *For example the AR(2) process:*

$$Y_t = 0.3Y_{t-1} + 0.8Y_{t-2} + a_t \tag{3.53}$$

is not *stationary since*

$$\phi_1 + \phi_2 = 0.3 + 0.8 = 1.1 > 1. \tag{3.54}$$

The same argument can be applied to $\phi(-1)$ to obtain:

Theorem 89 *A necessary condition for the stationarity of an AR(p) process is that* $\phi(-1) > 0$ *or:*

$$\sum_{j=1}^{p} (-1)^j \phi_j < 1. \tag{3.55}$$

Proof. Same as the proof of Theorem 87 except all arguments use $(1 + r_i)$ instead of $(1 - r_i)$. ■

3.4.2 Stationarity of an AR(2)

Suppose we apply the necessary conditions (3.49), (3.50) and (3.55) to the AR(2):

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + a_t. \tag{3.56}$$

For the AR(2) these also turn out to be sufficient for stationarity. That is:

Theorem 90 *For the AR(2) model in* (3.56) *a necessary and sufficient condition for stationarity is:*

$$\begin{aligned} |\phi_2| &< 1 \\ \phi_1 + \phi_2 &< 1 \\ \phi_2 - \phi_1 &< 1. \end{aligned} \tag{3.57}$$

Proof. We have already shown these conditions are necessary. To prove sufficiency note that from:

$$1 - \phi_1 B - \phi_2 B^2 = (1 - r_1 B)(1 - r_2 B)$$

we have:

$$\begin{aligned} r_1 + r_2 &= \phi_1 \\ r_1 r_2 &= -\phi_2. \end{aligned}$$

From $|\phi_2| = |r_1 r_2| = |r_1| |r_2| < 1$ it follows that either $|r_1| < 1$ or $|r_2| < 1$. If r_1 and r_2 are complex then $|r_1| = |r_2| < 1$ and stationarity follows. Therefore we need only consider the case where the roots are real. Assume without loss of generality that $|r_2| < 1$ so it remains to be shown that $|r_1| < 1$. Now from $\phi_1 + \phi_2 < 1$ we conclude that:

$$\phi_1 + \phi_2 = r_1 + r_2 - r_1 r_2 < 1$$

or:

$$r_1 (1 - r_2) < 1 - r_2.$$

Since $(1 - r_2) > 0$ if follows that $r_1 < 1$. Now from $\phi_2 - \phi_1 < 1$ we conclude that:

$$\phi_2 - \phi_1 = -r_1 r_2 - r_1 - r_2 < 1$$

or:

$$-r_1 (1 + r_2) < (1 + r_2).$$

Since $(1 - r_2) > 0$ if follows that $r_1 > -1$. Therefore $|r_1| < 1$ and $|r_2| < 1$ and so the process is stationary. ■

Remark 91 *Note that if the roots r_1 and r_2 are complex, which occurs if and only if $\phi_1^2 + 4\phi_2 < 0$, then a necessary and sufficient condition for stationarity is $|\phi_2| < 1$; that is we do not need the other two conditions in* (3.57).

3.5 The Autocorrelation Function

Consider now the problem of calculating the autocovariance function $\gamma(k)$ and the autocorrelation function $\rho(k)$ for an AR(p) process. We will derive $\gamma(0)$ and $\rho(k)$ for $k > 1$ since given these, $\gamma(k)$ can be calculated as

$$\gamma(k) = \rho(k)\gamma(0). \tag{3.58}$$

First using Theorem 56 we have:

$$E[Y_t a_t] = \psi_o \sigma^2 = \sigma^2 \tag{3.59}$$

since $\psi_o = 1$.

Multiplying both sides of

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + a_t$$

by Y_t and taking expectations yields:

$$\gamma(0) = E[Y_t^2] = \phi_1 \underbrace{E[Y_{t-1}Y_t]}_{\equiv\gamma(1)} + \phi_2 \underbrace{E[Y_{t-2}Y_t]}_{\equiv\gamma(2)} + \cdots + \phi_p \underbrace{E[Y_{t-p}Y_t]}_{\equiv\gamma(p)} + \underbrace{E[a_t Y_t]}_{\sigma^2} \tag{3.60}$$

so that

$$\gamma(0) = \phi_1\gamma(1) + \phi_2\gamma(2) + \cdots + \phi_p\gamma(p) + \sigma^2. \tag{3.61}$$

Now since $\gamma(k) = \rho(k)\gamma(0)$ we have solving for $\gamma(0)$ the following theorem:

Theorem 92 *For an AR(p) process:*

$$\gamma(0) = \frac{\sigma^2}{1 - \phi_1\rho(1) - \phi_2\rho(2) - \cdots - \phi_p\rho(p)}.$$

Thus given $\phi_1, \phi_2, \ldots \phi_p, \sigma^2$ and $\rho(1), \rho(2), \ldots \rho(p)$ we can calculate $\gamma(0)$.

Example 93 *For $Y_t = \phi Y_{t-1} + a_t$ we have: $\rho(1) = \phi$ so that:*

$$\begin{aligned} \gamma(0) &= \frac{\sigma^2}{1 - \phi\rho(1)} \\ &= \frac{\sigma^2}{1 - \phi^2} \end{aligned} \tag{3.62}$$

a formula we have already derived using other methods.

The autocorrelation function $\rho(k)$ can then be recursively calculated using:

Theorem 94 *For an AR(p) process:*

$$\rho(k) = \phi_1 \rho(k-1) + \phi_2 \rho(k-2) + \cdots + \phi_p \rho(k-p)$$

with starting values determined by

$$\rho(0) = 1 \text{ and } \rho(-k) = \rho(k).$$

Proof. Multiply both sides of

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + a_t \tag{3.63}$$

by Y_{t-k} (for $k > 0$) and take expectations. Using the fact that $E[a_t Y_{t-k}] = 0$ it follows that:

$$\underbrace{E[Y_t Y_{t-k}]}_{\equiv\gamma(k)} = \phi_1 \underbrace{E[Y_{t-1} Y_{t-k}]}_{\equiv\gamma(k-1)} + \phi_2 \underbrace{E[Y_{t-2} Y_{t-k}]}_{\equiv\gamma(k-2)} + \cdots + \phi_p \underbrace{E[Y_{t-p} Y_{t-k}]}_{\equiv\gamma(k-p)} \tag{3.64}$$

and dividing both sides by $\gamma(0)$ and using $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$ the theorem follows. ■

Thus $\rho(k)$ follows the same linear p^{th} order difference equation as ψ_k except that the starting values are different. Thus following the proof of Theorem 76 it follows that $\rho(k)$ will have the same solution except that the weights on r_j^k will be different. We thus have:

Theorem 95 *For an AR(p) process:*

$$\rho(k) = B_1 r_1^k + B_2 r_2^k + \cdots + B_p r_p^k$$

where $\phi(r_i^{-1}) = 0$, $\rho(0) = 1$ *and* $\rho(-k) = \rho(k)$.

We thus conclude for a stationary AR(p) that $\rho(k)$ has the short-memory property given in Section 2.2. In particular:

Theorem 96 *For a stationary AR(p) process* $\rho(k) = O\left(\tau^k\right)$ *or it has the short-memory property:*

$$|\rho(k)| \leq B\tau^k$$

where

$$0 \leq \tau = \max_i \left[|r_i|\right] < 1.$$

Proof. Same as the proof of Theorem 83. ■

3.6 Forecasting

Consider the problem of forecasting Y_{t+k} given the information set at time t. We have:

$$E_t[Y_{t+k}] = \phi_1 E_t[Y_{t+k-1}] + \phi_2 E_t[Y_{t+k-2}] + \cdots + \phi_p E_t[Y_{t+k-p}] + E_t[a_{t+k}]. \tag{3.65}$$

Since $E_t[a_{t+k}] = 0$ for $k > 0$ it follows that:

Theorem 97 *For an AR(p) process:*

$$E_t[Y_{t+k}] = \phi_1 E_t[Y_{t+k-1}] + \phi_2 E_t[Y_{t+k-2}] + \cdots + \phi_p E_t[Y_{t+k-p}] \tag{3.66}$$

where:

$$E_t[Y_{t+k}] = Y_{t+k} \quad \textit{for} \quad k \leq 0. \tag{3.67}$$

Thus the forecasts follow the same p^{th} order difference equation as ψ_k and $\rho(k)$ except that the starting values in (3.67) are different. These starting values come from the fact that forecasts of the past or present are just the already known values. We therefore have immediately that:

Theorem 98 *For an AR(p) process:*

$$E_t[Y_{t+k}] = C_{1t} r_1^k + C_{2t} r_2^k + \cdots + C_{pt} r_p^k \tag{3.68}$$

where r_i are the roots given by $\phi(r_i^{-1}) = 0$ and

$$E_t[Y_{t+k}] = Y_{t+k} \quad \textit{for} \quad k \leq 0.$$

Remark 99 *Unlike A_j and B_j for ψ_k and $\rho(k)$, there is a t subscript on C_{jt}. This is because the starting values depend on the information set at time t which depends on t.*

It follows that $E_t[Y_{t+k}]$, like ψ_k and $\rho(k)$ has the short-memory property and so:

Theorem 100 *For a stationary AR(p) process $E_t[Y_{t+k}] = O(\tau^k)$ or it has the short-memory property:*

$$|E_t[Y_{t+k}]| \leq C_t \tau^k$$

where

$$0 \leq \tau = \max_i [|r_i|] < 1.$$

Proof. Same as the proof of Theorem 83. ■

3.7 Some Worked Examples

Let us now consider a number of worked examples with AR(2) processes.

3.7.1 A Stationary AR(2) with Real Roots

Checking for Stationarity

Consider the AR(2) process:

$$Y_t = 0.3Y_{t-1} + 0.4Y_{t-2} + a_t, \ , \ \sigma = 0.01. \tag{3.69}$$

To determine stationarity we first work directly with $\phi\left(B\right)$ here given by:

$$\phi(B) = 1 - 0.3B - 0.4B^2. \tag{3.70}$$

Calculating the roots of $\phi(B) = 0$ we find that:

$$1 - 0.3B - 0.4B^2 = 0 \Rightarrow B = \frac{0.3 \pm \sqrt{(-0.3)^2 - 4(1)(-0.4)}}{2(-0.4)} \tag{3.71}$$

or

$$B_1 = -2 \text{ and } B_2 = 1.2. \tag{3.72}$$

Since $|B_1| > 1$ and $|B_2| > 1$ we conclude that the process is stationary.

We can calculate the roots of $\tilde{\phi}(r) = 0$ where:

$$\tilde{\phi}(r) = r^2 - 0.3r - 0.4 \tag{3.73}$$

which yields:

$$r = \frac{0.3 \pm \sqrt{(0.3)^2 - 4(-0.4)}}{2} \tag{3.74}$$

or

$$r_1 = \frac{4}{5} \text{ and } r_2 = -\frac{1}{2}. \tag{3.75}$$

Since $|r_1| < 1$ and $|r_2| < 1$ we again can conclude again that the series is stationary.

Finally we can check stationarity from (3.57) . Since:

$$\begin{aligned} |\phi_2| &= 0.4 < 1 \\ \phi_1 + \phi_2 &= 0.7 < 1 \\ \phi_2 - \phi_1 &= 0.1 < 1 \end{aligned} \tag{3.76}$$

it follows that the process is stationary.

Infinite Moving Average Weights

Recursive Calculations To calculate the infinite moving average weights recursively we use:

$$\begin{aligned} \psi_k &= 0.3\psi_{k-1} + 0.4\psi_{k-2} \\ \psi_0 &= 1, \;\; \psi_{-1} = 0 \end{aligned} \tag{3.77}$$

so that:

$$\psi_1 = 0.3 \underbrace{\psi_0}_{=1} + 0.4 \underbrace{\psi_{-1}}_{=0} = 0.3 \tag{3.78}$$

$$\psi_2 = 0.3 \underbrace{\psi_1}_{=0.3} + 0.4 \underbrace{\psi_0}_{=1} = 0.49$$

$$\psi_3 = 0.3 \underbrace{\psi_2}_{=0.49} + 0.4 \underbrace{\psi_1}_{=0.3} = 0.267$$

$$\begin{aligned} \psi_4 &= 0.3 \underbrace{\psi_3}_{=0.267} + 0.4 \underbrace{\psi_2}_{=0.49} = 0.276, \\ & \text{etc..} \end{aligned}$$

Solving the Difference Equation

We can also directly solve the above difference equation. The roots of

$$r^2 - 0.3r - 0.4 = 0 \tag{3.79}$$

are given by: $r_1 = \frac{4}{5}$ and $r_2 = -\frac{1}{2}$ so that:

$$\psi_k = A_1 \left(\frac{4}{5}\right)^k + A_2 \left(-\frac{1}{2}\right)^k . \tag{3.80}$$

To find A_1 and A_2 use the fact that $\psi_0 = 1$ and $\psi_{-1} = 0$ so that:

$$\begin{aligned} \psi_{-1} &= 0 = A_1 \left(\frac{4}{5}\right)^{-1} + A_2 \left(-\frac{1}{2}\right)^{-1} \\ \psi_0 &= 1 = A_1 \left(\frac{4}{5}\right)^0 + A_2 \left(-\frac{1}{2}\right)^0 \end{aligned} \tag{3.81}$$

or in matrix notation:

$$\begin{bmatrix} \frac{5}{4} & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} . \tag{3.82}$$

Solving for A_1 and A_2 we find that:

$$\begin{aligned} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} &= \begin{bmatrix} \frac{5}{4} & -2 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{8}{13} \\ \frac{5}{13} \end{bmatrix} \end{aligned} \tag{3.83}$$

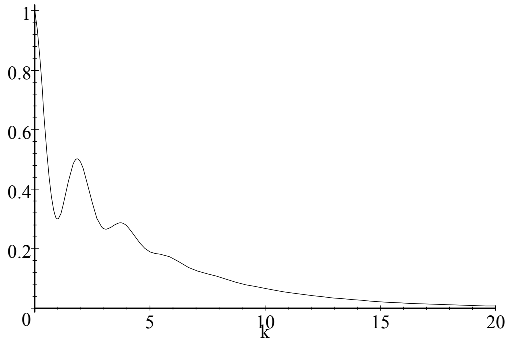
so that:

$$\psi_k = \frac{8}{13}\left(\frac{4}{5}\right)^k + \frac{5}{13}\left(-\frac{1}{2}\right)^k. \tag{3.84}$$

You can verify that this is the correct solution by substituting in values of k. For example for $k = 3$ we have:

$$\begin{aligned} \psi_3 &= \frac{8}{13}\left(\frac{4}{5}\right)^3 + \frac{5}{13}\left(-\frac{1}{2}\right)^3 \\ &= 0.267 \end{aligned} \tag{3.85}$$

which is identical to what we found with the recursive calculation above. A (connected) plot of ψ_k for $k = 1, 2, \ldots 20$ is given below:



Plot of ψ_k

Note the exponential decay reflecting the short-memory property of $\psi_k = O\left(\tau^k\right)$.

Autocorrelations

Recursive Calculations

For the autocorrelation function we have:

$$\rho(k) = 0.3\rho(k-1) + 0.4\rho(k-2). \tag{3.86}$$

We can calculate this recursively as:

$$\begin{aligned} \rho(0) &= 1 \qquad (3.87) \\ \rho(1) &= 0.3\underbrace{\rho(0)}_{=1} + 0.4\underbrace{\rho(-1)}_{\rho(1)} \end{aligned}$$

so that

$$\rho(1) = 0.3 + 0.4\rho(1) \qquad (3.88)$$

or

$$\rho(1) = \frac{0.3}{0.6} = 0.5. \qquad (3.89)$$

Continuing we have:

$$\begin{aligned} \rho(2) &= 0.3\underbrace{\rho(1)}_{=0.5} + 0.4\underbrace{\rho(0)}_{=1} = 0.55 \qquad (3.90) \\ \rho(3) &= 0.3\underbrace{\rho(2)}_{=0.55} + 0.4\underbrace{\rho(1)}_{=0.5} = 0.365 \\ \rho(4) &= 0.3\underbrace{\rho(3)}_{=0.365} + 0.4\underbrace{\rho(2)}_{=0.55} = 0.3295 \\ &\text{etc..} \end{aligned}$$

Given $\rho(1) = 0.5$, $\rho(2) = 0.55$ we can determine $\gamma(0)$ from Theorem 92 as:

$$\begin{aligned} \gamma(0) &= \frac{\sigma^2}{1 - \phi_1\rho(1) - \phi_2\rho(2)} \qquad (3.91) \\ &= \frac{(0.01)^2}{1 - (0.3)(0.5) - (0.4)(0.55)} \\ &= 0.000159 \end{aligned}$$

so that the unconditional standard deviation is:

$$\gamma(0)^{\frac{1}{2}} = \sqrt{0.000159} = 0.0126. \qquad (3.92)$$

Solving the Difference Equation To solve the difference equation for $\rho(k)$ we have:

$$\rho(k) = B_1\left(\frac{4}{5}\right)^k + B_2\left(-\frac{1}{2}\right)^k \qquad (3.93)$$

with starting values $\rho(0) = 1$ and $\rho(-1) = \rho(1) = \frac{1}{2}$ so that:

$$\begin{bmatrix} \frac{5}{4} & -2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$$

or:

$$\begin{aligned}\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} &= \begin{bmatrix} \frac{5}{4} & -2 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{10}{13} \\ \frac{3}{13} \end{bmatrix}\end{aligned} \tag{3.94}$$

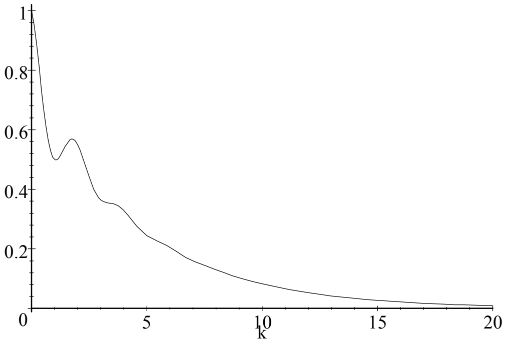
and hence:

$$\rho(k) = \frac{10}{13}\left(\frac{4}{5}\right)^k + \frac{3}{13}\left(-\frac{1}{2}\right)^k. \tag{3.95}$$

We can again verify that this is correct for say $k = 2$ since:

$$\begin{aligned}\rho(2) &= \frac{10}{13}\left(\frac{4}{5}\right)^2 + \frac{3}{13}\left(-\frac{1}{2}\right)^2 \\ &= 0.55\end{aligned} \tag{3.96}$$

which is identical to the recursive calculation result. A (connected) plot of $\rho\,(k)$ is given below:



Plot of $\rho\,(k)$

Note the exponential decay due to the short-memory property.

Forecasts

Recursive Calculations To calculate forecasts from this model suppose that:

$$Y_{t-1} = 0.01,\ Y_t = 0.02,\ \sigma = 0.01. \tag{3.97}$$

Then since $E_t[Y_{t-1}] = 0.01$ and $E_t[Y_t] = 0.02$ we have:

$$\begin{aligned}E_t[Y_{t+1}] &= 0.3\underbrace{E_t[Y_t]}_{=0.02} + 0.4\underbrace{E_t[Y_{t-1}]}_{=0.01} = 0.01 \\ E_t[Y_{t+2}] &= 0.3\underbrace{E_t[Y_{t+1}]}_{=0.01} + 0.4\underbrace{E_t[Y_t]}_{=0.02} = 0.011\end{aligned} \tag{3.98}$$

$$\begin{aligned} E_t[Y_{t+3}] &= 0.3\underbrace{E_t[Y_{t+2}]}_{=0.011} + 0.4\underbrace{E_t[Y_{t+1}]}_{=0.01} = 0.0073 \\ E_t[Y_{t+4}] &= 0.3\underbrace{E_t[Y_{t+3}]}_{=0.0073} + 0.4\underbrace{E_t[Y_{t+2}]}_{=0.01} = 0.00619. \end{aligned}$$

Solving the Difference Equation We can also solve the difference equation. Since

$$E_t[Y_{t+k}] = C_{1t}r_1^k + C_{2t}r_2^k \tag{3.99}$$

with $r_1 = \frac{4}{5}$ and $r_2 = -\frac{1}{2}$ we have:

$$E_t[Y_{t+k}] = C_{1t}\left(\frac{4}{5}\right)^k + C_{2t}\left(-\frac{1}{2}\right)^k. \tag{3.100}$$

We calculate C_{1t} and C_{2t} using the starting values $E[Y_{t-1}] = 0.01$ and $E_t[Y_t] = 0.02$ so that:

$$\begin{aligned} C_{1t}\left(\frac{4}{5}\right)^{-1} + C_{2t}\left(-\frac{1}{2}\right)^{-1} &= E[Y_{t-1}] = 0.01 \\ C_{1t}\left(\frac{4}{5}\right)^{0} + C_{2t}\left(-\frac{1}{2}\right)^{0} &= E[Y_t] = 0.02 \end{aligned}$$

or in matrix notation:

$$\begin{bmatrix} \frac{5}{4} & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} C_{1t} \\ C_{2t} \end{bmatrix} = \begin{bmatrix} 0.01 \\ 0.02 \end{bmatrix} \tag{3.101}$$

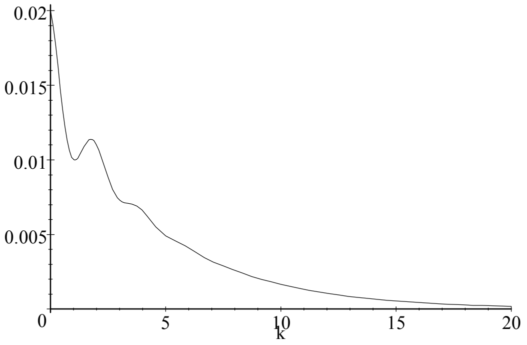
and hence:

$$\begin{aligned} \begin{bmatrix} C_{1t} \\ C_{2t} \end{bmatrix} &= \begin{bmatrix} \frac{5}{4} & -2 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.01 \\ 0.02 \end{bmatrix} \\ &= \begin{bmatrix} 0.0154 \\ 0.0046 \end{bmatrix}. \end{aligned} \tag{3.102}$$

We therefore have:

$$E_t[Y_{t+k}] = 0.0154\left(\frac{4}{5}\right)^k + 0.0046\left(-\frac{1}{2}\right)^k. \tag{3.103}$$

A (connected) plot of $E_t[Y_{t+k}]$ is given below:



A plot of $E_t\left[Y_{t+k}\right]$ as a function of k

Note the exponential decay indicative of the short-memory property of $E_t\left[Y_{t+k}\right] = O\left(\tau^k\right)$.

Confidence Intervals for Forecasts

To calculate confidence intervals for the forecasts we use the fact that for any stationary time series we have from the Wold representation and Theorem 61 that:

$$Var_t[Y_{t+k}] = \sigma^2(1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{k-1}^2) \tag{3.104}$$

where the ψ_k's are calculated recursively from:

$$\begin{aligned} \psi_k &= 1.4\psi_{k-1} - 0.7\psi_{k-2} \\ \psi_0 &= 1, \;\; \psi_{-1} = 0. \end{aligned} \tag{3.105}$$

Recall that $\psi_1 = 1.4$, $\psi_2 = 1.26$, $\psi_3 = 0.784$ and $\psi_4 = 0.2156$. Thus:

$$\begin{aligned} Var_t[Y_{t+1}] &= \sigma^2 = (0.01)^2 \\ Var_t[Y_{t+1}] &= \sigma^2(1 + \psi_1^2) = (0.01)^2\left(1 + 1.4^2\right) \\ Var_t[Y_{t+2}] &= \sigma^2(1 + \psi_1^2 + \psi_2^2) = (0.01)^2\left(1 + 1.4^2 + 1.26^2\right) \\ Var_t[Y_{t+3}] &= \sigma^2(1 + \psi_1^2 + \psi_2^2 + \psi_3^2) = (0.01)^2\left(1 + 1.4^2 + 1.26^2 + 0.784^2\right) \\ & \quad etc.. \end{aligned}$$

A 95% confidence interval for Y_{t+k} is then given by:

$$E_t[Y_{t+k}] \pm 1.96\sqrt{Var_t[Y_{t+k}]}. \tag{3.106}$$

As $k \to \infty$ $E_t[Y_{t+k}] \to E\left[Y_t\right] = 0$ and $Var_t[Y_{t+k}] \to \gamma(0)$ and the confidence interval would be

$$0 \pm 1.96\sqrt{\gamma(0)}. \tag{3.107}$$

For this model we have from (3.92) that $\gamma(0) = 0.025$. Thus we have:

Forecast Confidence Intervals

$k = 1$	0.021 ± 0.0196
$k = 2$	0.0154 ± 0.037
$k = 3$	0.007 ± 0.0418
$k = 4$	-0.001 ± 0.0445
$k = 5$	-0.006 ± 0.0447
$k = \infty$	0 ± 0.049

3.7.2 A Nonstationary AR(2) with Real Roots

Let us consider the $AR(2)$ process:

$$Y_t = 0.7Y_{t-1} + 0.6Y_{t-2} + a_t. \tag{3.108}$$

Here we have:

$$\tilde{\phi}(r) = r^2 - 0.7r - 0.6 \tag{3.109}$$

which yields:

$$r_1 = 1.2 \text{ and } r_2 = -\frac{1}{2}. \tag{3.110}$$

Since $|r_1| = 1.2 > 1$ we conclude that the series is nonstationary.

Alternatively we would conclude from the fact that:

$$\phi_1 + \phi_2 = 0.7 + 0.6 = 1.3 > 1$$

which violates (3.57) and so the process is nonstationary.

Infinite Moving Average Weights

Recursive Calculations To calculate the infinite moving average weights recursively we use:

$$\begin{aligned} \psi_k &= 0.7\psi_{k-1} + 0.6\psi_{k-2} \\ \psi_0 &= 1,\ \psi_{-1} = 0 \end{aligned} \tag{3.111}$$

so that:

$$\psi_1 = 0.7\underbrace{\psi_0}_{=1} + 0.6\underbrace{\psi_{-1}}_{=0} = 0.7 \tag{3.112}$$

$$\psi_2 = 0.7\underbrace{\psi_1}_{=0.7} + 0.6\underbrace{\psi_0}_{=1} = 1.09$$

$$\psi_3 = 0.7\underbrace{\psi_2}_{=1.09} + 0.6\underbrace{\psi_1}_{=0.7} = 1.183$$

$$\begin{aligned}\psi_4 &= 0.7\underbrace{\psi_3}_{=1.183} + 0.6\underbrace{\psi_2}_{=1.09} = 1.482, \\ &\text{etc..}\end{aligned}$$

Note that the weights are getting bigger and bigger, which is a reflection of the fact that the process is not stationary.

Solving the Difference Equation We can also directly solve the above difference equation. The roots of

$$r^2 - 0.7r - 0.6 = 0 \tag{3.113}$$

are given by: $r_1 = \frac{6}{5}$ and $r_2 = -\frac{1}{2}$ so that:

$$\psi_k = A_1\left(\frac{6}{5}\right)^k + A_2\left(-\frac{1}{2}\right)^k. \tag{3.114}$$

To find A_1 and A_2 use the fact that $\psi_0 = 1$ and $\psi_{-1} = 0$ so that:

$$\begin{aligned}\psi_{-1} &= 0 = A_1\left(\frac{6}{5}\right)^{-1} + A_2\left(-\frac{1}{2}\right)^{-1} \\ \psi_0 &= 1 = A_1\left(\frac{6}{5}\right)^0 + A_2\left(-\frac{1}{2}\right)^0\end{aligned} \tag{3.115}$$

or in matrix notation:

$$\begin{bmatrix} \frac{5}{6} & -2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{3.116}$$

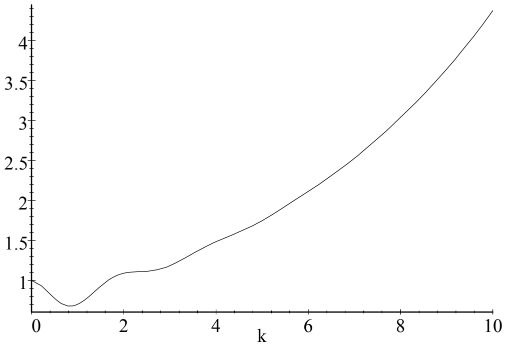
Solving for A_1 and A_2 we find that:

$$\begin{aligned}\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} &= \begin{bmatrix} \frac{5}{6} & -2 \\ 1 & 1 \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{12}{17} \\ \frac{5}{17} \end{bmatrix}\end{aligned} \tag{3.117}$$

so that:

$$\psi_k = \frac{12}{17}\left(\frac{6}{5}\right)^k + \frac{5}{17}\left(-\frac{1}{2}\right)^k. \tag{3.118}$$

We can see now why $\psi_k' s$ do not converge to zero since the term involving $\left(\frac{6}{5}\right)^k$ will diverge as $k \to \infty$ even though the term involving $\left(-\frac{1}{2}\right)^k$ will converge to zero. A (connected) plot of ψ_k for $k = 1, 2, \ldots 10$ is given below shows this:



A plot of ψ_k.

Autocorrelations

Recursive Calculations For the autocorrelation function we have:

$$\rho(k) = 0.7\rho(k-1) + 0.6\rho(k-2). \tag{3.119}$$

Since the series is not stationary it strictly speaking does not have an autocorrelation function. We can nevertheless calculate the $\rho\left(k\right)' s$ recursively and see what happens. We obtain:

$$\begin{aligned} \rho(0) &= 1 \qquad (3.120) \\ \rho(1) &= 0.7\underbrace{\rho(0)}_{=1} + 0.6\underbrace{\rho(-1)}_{\rho(1)} \end{aligned}$$

so that

$$\rho(1) = 0.7 + 0.6\rho(1) \tag{3.121}$$

or

$$\rho(1) = \frac{0.7}{0.4} = \frac{7}{4} > 1 \tag{3.122}$$

which is impossible since $-1 \leq \rho\left(k\right) \leq 1$. Again we conclude that the series must be nonstationary.

3.7.3 A Nonstationary AR(2) with a Unit Root

Let us consider the $AR(2)$ process:

$$Y_t = 0.7Y_{t-1} + 0.3Y_{t-2} + a_t. \tag{3.123}$$

Here we have:

$$\tilde{\phi}(r) = r^2 - 0.7r - 0.3 \tag{3.124}$$

which yields:

$$r_1 = 1.0 \text{ and } r_2 = -0.3. \tag{3.125}$$

Since $r_1 = 1.0$ which is not less than 1 in absolute value, we conclude that the series is nonstationary. We can also see this by the fact that:

$$\phi_1 + \phi_2 = 0.7 + 0.3 = 1.0 = 1$$

which violates (3.57).

This series has a unit root, and is in fact on the border between stationarity and nonstationarity.

Infinite Moving Average Weights

Recursive Calculations To calculate the infinite moving average weights recursively we use:

$$\begin{aligned} \psi_k &= 0.7\psi_{k-1} + 0.3\psi_{k-2} \\ \psi_0 &= 1, \ \psi_{-1} = 0 \end{aligned} \tag{3.126}$$

so that:

$$\psi_1 = 0.7 \underbrace{\psi_0}_{=1} + 0.3 \underbrace{\psi_{-1}}_{=0} = 0.7 \tag{3.127}$$

$$\psi_2 = 0.7 \underbrace{\psi_1}_{=0.7} + 0.3 \underbrace{\psi_0}_{=1} = 0.79$$

$$\psi_3 = 0.7 \underbrace{\psi_2}_{=0.79} + 0.3 \underbrace{\psi_1}_{=0.7} = 0.763$$

$$\begin{aligned} \psi_4 &= 0.7 \underbrace{\psi_3}_{=0.763} + 0.3 \underbrace{\psi_2}_{=0.79} = 0.7711, \\ & \text{etc..} \end{aligned}$$

As we shall see, the ψ_k *'s* are not diverging, as with the previous nonstationary example, nor are they converging to 0 as required by stationary, but they are converging to a non-zero constant.

Solving the Difference Equation We can also directly solve the above difference equation. The roots of

$$r^2 - 0.7r - 0.3 = 0 \tag{3.128}$$

are given by: $r_1 = 1$ and $r_2 = -0.3$ so that:

$$\psi_k = A_1 (1)^k + A_2 (-0.3)^k . \tag{3.129}$$

To find A_1 and A_2 use the fact that $\psi_0 = 1$ and $\psi_{-1} = 0$ so that:

$$\begin{aligned} \psi_{-1} &= 0 = A_1 1^{-1} + A_2 (-0.3)^{-1} \qquad (3.130)\\ \psi_0 &= 1 = A_1 1^0 + A_2 (-0.3)^0 \end{aligned}$$

or in matrix notation:

$$\begin{bmatrix} 1 & -\frac{1}{0.3} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} . \tag{3.131}$$

Solving for A_1 and A_2 we find that:

$$\begin{aligned} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} &= \begin{bmatrix} 1 & -\frac{1}{0.3} \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad (3.132)\\ &= \begin{bmatrix} 0.76923 \\ 0.23077 \end{bmatrix} \end{aligned}$$

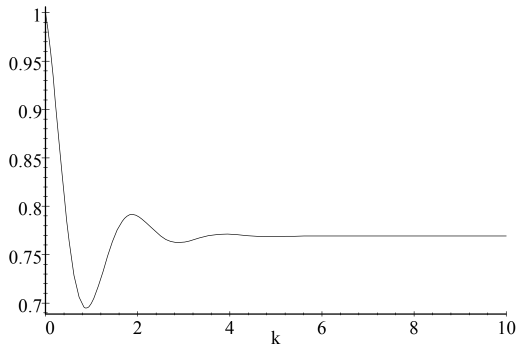
so that:

$$\begin{aligned} \psi_k &= 0.76923 (1)^k + 0.23077 (-0.3)^k \qquad (3.133)\\ &= 0.76923 + 0.23077 (-0.3)^k \qquad (3.134) \end{aligned}$$

since $1^k = 1$ for all k. Thus we see that as $k \to \infty$

$$\psi_k \to 0.76923$$

as illustrated in the (connected) plot of ψ_k below:



.

A plot of ψ_k.

This means that shocks: a_{t-k} will have a permanent effect on Y_t.

In general it can be shown that if an AR(p) has a unit root so that:

$$\phi(B) = (1-B)\tilde{\phi}(B)$$

then as $k \to \infty$

$$\psi_k \to \frac{1}{\tilde{\phi}(1)}.$$

In the above example:

$$\begin{aligned}\phi(B) &= 1-0.7B-0.3B^2 \\ &= (1-B)(1+0.3B)\end{aligned}$$

so that $\tilde{\phi}(B) = (1+0.3B)$ and so

$$\psi_k \to \frac{1}{\tilde{\phi}(1)} = \frac{1}{(1+0.3\times 1)} = 0.76923.$$

3.7.4 A Stationary AR(2) with Complex Roots

Let us consider the $AR(2)$ process:

$$Y_t = \frac{3}{2}Y_{t-1} - \frac{5}{8}Y_{t-2} + a_t,\ \sigma = 0.01. \tag{3.135}$$

Calculating the roots of $\tilde{\phi}(r) = 0$ where:

$$\tilde{\phi}(r) = r^2 - \frac{3}{2}r + \frac{5}{8} \tag{3.136}$$

yields:

$$r = \frac{\frac{3}{2} \pm \sqrt{(\frac{3}{2})^2 - 4(\frac{5}{8})}}{2} \tag{3.137}$$

or

$$r_1 = \frac{3}{4} + \frac{1}{4}i \text{ and } r_2 = \frac{3}{4} - \frac{1}{4}i. \tag{3.138}$$

We conclude that the process is stationary since:

$$|r_1| = |r_2| = \sqrt{\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2} = \sqrt{\frac{5}{8}} < 1. \tag{3.139}$$

Alternatively since we have complex roots it is enough to verify that

$$|\phi_2| = \frac{5}{8} < 1$$

to be sure that the process is stationary.

Infinite Moving Average Weights

Recursive Calculations To calculate the infinite moving average weights recursively we use:

$$\begin{aligned} \psi_k &= \frac{3}{2}\psi_{k-1} - \frac{5}{8}\psi_{k-2} \\ \psi_0 &= 1, \;\; \psi_{-1} = 0 \end{aligned} \tag{3.140}$$

so that:

$$\begin{aligned} \psi_1 &= \frac{3}{2}\underbrace{\psi_0}_{=1} - \frac{5}{8}\underbrace{\psi_{-1}}_{=0} = \frac{3}{2} \\ \psi_2 &= \frac{3}{2}\underbrace{\psi_1}_{=\frac{3}{2}} - \frac{5}{8}\underbrace{\psi_0}_{=1} = 1.625 \\ \psi_3 &= \frac{3}{2}\underbrace{\psi_2}_{=1.625} - \frac{5}{8}\underbrace{\psi_1}_{=\frac{3}{2}} = 1.5 \\ \psi_4 &= \frac{3}{2}\underbrace{\psi_3}_{=1.5} - \frac{5}{8}\underbrace{\psi_2}_{=1.625} = 1.234, \\ \psi_5 &= \frac{3}{2}\underbrace{\psi_4}_{=1.234} - \frac{5}{8}\underbrace{\psi_3}_{=1.5} = 0.9135 \\ & \text{etc..} \end{aligned} \tag{3.141}$$

We can also directly solve the above difference equation. Using the roots:

$$r_1 = \frac{3}{4} + \frac{1}{4}i, \; r_2 = \frac{3}{4} - \frac{1}{4}i \tag{3.142}$$

we have:

$$\psi_k = A_1\left(\frac{3}{4} + \frac{1}{4}i\right)^k + A_2\left(\frac{3}{4} - \frac{1}{4}i\right)^k. \tag{3.143}$$

To find A_1 and A_2 use the fact that $\psi_0 = 1$ and $\psi_{-1} = 0$ so that:

$$\begin{aligned} \psi_{-1} &= 0 = A_1\left(\frac{3}{4} + \frac{1}{4}i\right)^{-1} + A_2\left(\frac{3}{4} - \frac{1}{4}i\right)^{-1} \\ \psi_0 &= 1 = A_1\left(\frac{3}{4} + \frac{1}{4}i\right)^{0} + A_2\left(\frac{3}{4} - \frac{1}{4}i\right)^{0} \end{aligned} \tag{3.144}$$

or in matrix notation:

$$\begin{bmatrix} \left(\frac{3}{4} + \frac{1}{4}i\right)^{-1} & \left(\frac{3}{4} - \frac{1}{4}i\right)^{-1} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{3.145}$$

Solving for A_1 and A_2 we find that:

$$\begin{aligned} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} &= \begin{bmatrix} \left(\frac{3}{4}+\frac{1}{4}i\right)^{-1} & \left(\frac{3}{4}-\frac{1}{4}i\right)^{-1} \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2}-\frac{3}{2}i \\ \frac{1}{2}+\frac{3}{2}i \end{bmatrix} \end{aligned} \tag{3.146}$$

so that:

$$\psi_k = \left(\frac{1}{2}-\frac{3}{2}i\right)\left(\frac{3}{4}+\frac{1}{4}i\right)^k + \left(\frac{1}{2}+\frac{3}{2}i\right)\left(\frac{3}{4}-\frac{1}{4}i\right)^k . \tag{3.147}$$

This solution may not seem correct since we know that ψ_k is real while the solution seems to be complex. If you review your complex variable theory you will see that the two terms are complex conjugates of each other and so the imaginary parts will cancel. In fact using the polar representation of the complex numbers where:

$$\begin{aligned} \frac{1}{2}-\frac{3}{2}i &= \sqrt{\frac{5}{2}}e^{-\phi i} \\ \frac{1}{2}+\frac{3}{2}i &= \sqrt{\frac{5}{2}}e^{\phi i} \\ \frac{3}{4}+\frac{1}{4}i &= \sqrt{\frac{5}{8}}e^{\theta i} \\ \frac{3}{4}-\frac{1}{4}i &= \sqrt{\frac{5}{8}}e^{-\theta i}, \end{aligned} \tag{3.148}$$

and where:

$$\begin{aligned} \theta &= \arctan\left(\frac{1}{3}\right) = 0.321 \\ \phi &= \arctan(3) = 1.249 \end{aligned} \tag{3.149}$$

we have using Euler's theorem:

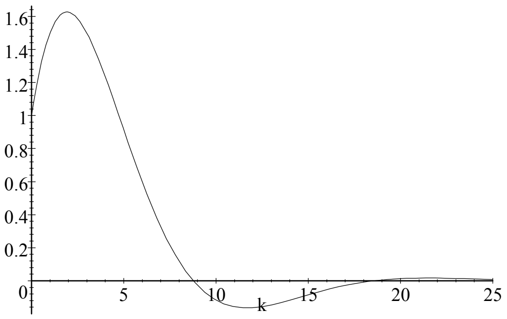
$$e^{ix} = \cos(x) + i\sin(x) \tag{3.150}$$

that:

$$\psi_k = \left(\frac{5}{8}\right)^{k/2}\left(\cos(k\theta) + 3\sin(k\theta)\right) \tag{3.151}$$

$$= \sqrt{10}\left(\frac{5}{8}\right)^{k/2}\sin(\theta(k+1)). \tag{3.152}$$

Thus ψ_k is an exponentially damped sine wave as the diagram plot below shows:



Plot of ψ_k

The fact that ψ_k is exponentially damped reflects the short-memory property.

3.7.5 A Nonstationary AR(2) with Complex Roots

Let us consider the $AR(2)$ process:

$$Y_t = \frac{3}{2}Y_{t-1} - \frac{9}{8}Y_{t-2} + a_t. \tag{3.153}$$

Working directly with: $\phi(B)$ we find that:

$$\phi(B) = 1 - \frac{3}{2}B + \frac{9}{8}B^2 \tag{3.154}$$

so that calculating the roots of $\phi(B) = 0$ we find that:

$$1 - \frac{3}{2}B + \frac{9}{8}B^2 = 0 \Rightarrow B = \frac{\frac{3}{2} \pm \sqrt{(-\frac{3}{2})^2 - 4(1)\left(\frac{9}{8}\right)}}{2(\frac{9}{8})} \tag{3.155}$$

or

$$B_1 = \frac{2}{3} + \frac{2}{3}i \text{ and } B_2 = \frac{2}{3} - \frac{2}{3}i. \tag{3.156}$$

where $i = \sqrt{-1}$.

Using this we see that:

$$|B_1| = |B_2| = \sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{2}{3}\right)^2} = \sqrt{\frac{8}{9}} < 1 \tag{3.157}$$

and so conclude that the process is *not* stationary.

Alternatively we can calculate the roots of $\tilde{\phi}(r) = 0$ where:

$$\tilde{\phi}(r) = r^2 - \frac{3}{2}r + \frac{9}{8} \tag{3.158}$$

which yields:

$$r = \frac{\frac{3}{2} \pm \sqrt{(\frac{3}{2})^2 - 4(\frac{9}{8})}}{2} \tag{3.159}$$

or

$$r_1 = \frac{3}{4} + \frac{3}{4}i \text{ and } r_2 = \frac{3}{4} - \frac{3}{4}i. \tag{3.160}$$

Using this we conclude that the process is not stationary since:

$$|r_1| = |r_2| = \sqrt{\left(\frac{3}{4}\right)^2 + \left(\frac{3}{4}\right)^2} = \sqrt{\frac{9}{8}} > 1. \tag{3.161}$$

Finally since the roots are complex it follows from (3.57) that since:

$$|\phi_2| = \frac{9}{8} > 1$$

that the process is nonstationary.

We can also directly solve the above difference equation. we have:

$$\psi_k = A_1 \left(\frac{3}{4} + \frac{3}{4}i\right)^k + A_2 \left(\frac{3}{4} - \frac{3}{4}i\right)^k. \tag{3.162}$$

To find A_1 and A_2 use the fact that $\psi_0 = 1$ and $\psi_{-1} = 0$ so that:

$$\begin{aligned} \psi_{-1} &= 0 = A_1 \left(\frac{3}{4} + \frac{3}{4}i\right)^{-1} + A_2 \left(\frac{3}{4} - \frac{3}{4}i\right)^{-1} \\ \psi_0 &= 1 = A_1 \left(\frac{3}{4} + \frac{3}{4}i\right)^{0} + A_2 \left(\frac{3}{4} - \frac{3}{4}i\right)^{0} \end{aligned} \tag{3.163}$$

or in matrix notation:

$$\begin{bmatrix} \left(\frac{3}{4} + \frac{3}{4}i\right)^{-1} & \left(\frac{3}{4} - \frac{3}{4}i\right)^{-1} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{3.164}$$

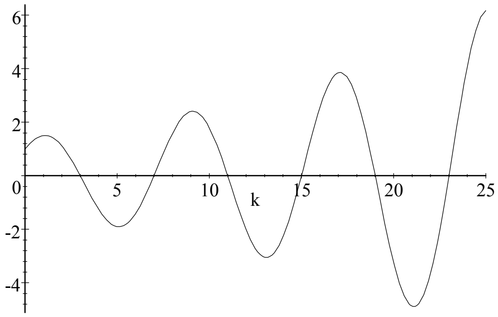
Solving for A_1 and A_2 we find that:

$$\begin{aligned} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} &= \begin{bmatrix} \left(\frac{3}{4} + \frac{3}{4}i\right)^{-1} & \left(\frac{3}{4} - \frac{3}{4}i\right)^{-1} \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2} - \frac{1}{2}i \\ \frac{1}{2} + \frac{1}{2}i \end{bmatrix} \end{aligned} \tag{3.165}$$

so that using Euler's theorem again we obtain:

$$\begin{aligned}\psi_k &= \left(\frac{1}{2}-\frac{1}{2}i\right)\left(\frac{3}{4}+\frac{3}{4}i\right)^k+\left(\frac{1}{2}+\frac{1}{2}i\right)\left(\frac{3}{4}-\frac{3}{4}i\right)^k \qquad (3.166)\\ &= \sqrt{2}\left(\sqrt{\frac{9}{8}}\right)^k\cos\left(\frac{\pi}{4}(k-1)\right).\end{aligned}$$

Note that ψ_k is an exponentially explosive sine wave with oscillations increasing in magnitude as k increases. This is illustrated in the diagram below:



Plot of ψ_k

3.8 Estimation and Hypothesis Testing

3.8.1 Least Squares Estimation

AR(1) Estimation

To begin first consider the problem of estimating an AR(1):

$$Y_t = \phi Y_{t-1} + a_t.$$

We can think of an AR(1) as a linear regression where:

$$Y_t = X_t\phi + a_t \qquad (3.167)$$

and where $X_t = Y_{t-1}$. Since the error term: a_t is uncorrelated with the regressor X_t it is natural to expect that the least squares estimator

$$\hat{\phi} = \frac{\sum_{t=1}^T X_t Y_t}{\sum_{t=1}^T X_t^2} = \frac{\sum_{t=1}^T Y_{t-1}Y_t}{\sum_{t=1}^T Y_{t-1}^2} \qquad (3.168)$$

will be well behaved, at least asymptotically. Thus subject to certain regularity conditions[2] we would expect that given a sample of T observations of Y_t that:

[2] Stationary AR(p) processes are ergodic, which means that sample means over one realization of the process converge to population means.

Proposition 101

$$\sqrt{T}(\hat{\phi}-\phi) \overset{a}{\sim} N\left[0, \sigma^2 (plim \frac{1}{T}\sum_{t=1}^{T} X_t^2)^{-1}\right]. \tag{3.169}$$

Since $X_t = Y_{t-1}$ we have given stationarity that

$$\begin{aligned} plim \frac{1}{T}\sum_{t=1}^{T} X_t^2 &= E\left[X_t^2\right] \\ &= E\left[Y_{t-1}^2\right] \\ &= \gamma(0) \\ &= \frac{\sigma^2}{1-\phi^2} \end{aligned}$$

so that:

$$\sqrt{T}(\hat{\phi}-\phi) \overset{a}{\sim} N\left[0, \sigma^2 (\frac{\sigma^2}{1-\phi^2})^{-1}\right]. \tag{3.170}$$

Note that σ^2 cancels from the asymptotic variance so that:

Theorem 102 *For a stationary AR(1) process:*

$$\sqrt{T}(\hat{\phi}-\phi) \overset{a}{\sim} N[0, 1-\phi^2].$$

Remark 103 *The fact that the asymptotic variance is independent of σ^2 reflects the fact that when σ^2 increases it increases both the variability of the error term in the regression, which reduces the accuracy of the $\hat{\phi}$, and it increases the variability of the regressor $X_t = Y_{t-1}$, which increases the accuracy of the $\hat{\phi}$. These two effects cancel each other out.*

AR(p) Estimation

The AR(p) process

$$Y_t = \sum_{j=1}^{p} Y_{t-j}\phi_j + a_t \tag{3.171}$$

can be written as a linear regression model as:

$$Y_t = X_t^T \phi + a_t \tag{3.172}$$

where:

$$\underset{px1}{X_t} = \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{bmatrix}, \quad \underset{px1}{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} \tag{3.173}$$

If we create a $T \times 1$ vector $Y = [Y_t]$ and a $T \times p$ matrix $X = \left[X_t^T\right]$ then the least squares estimator $\hat{\phi}$ has the usual formula:

$$\hat{\phi} = \left(X^T X\right)^{-1} X^T Y. \tag{3.174}$$

We have:

Theorem 104 *If Y_t is stationary the OLS estimator $\hat{\phi}$ given either by* (3.174) *or by:*

$$\hat{\phi} = \left(\sum_{t=1}^{T} X_t X_t^T\right)^{-1} \sum_{t=1}^{T} X_t Y_t \tag{3.175}$$

has an asymptotic distribution:

$$\sqrt{T}(\hat{\phi} - \phi) \overset{a}{\sim} N\left[0, \sigma^2 (\ plim \frac{1}{T} \sum_{t=1}^{T} X_t X_t^T)^{-1}\right]. \tag{3.176}$$

As with the AR(1) model the asymptotic variance is independent of σ^2 as:

Theorem 105 *For a stationary AR(p) process:*

$$\sqrt{T}(\hat{\phi} - \phi) \overset{a}{\sim} N\left[0, \delta R^{-1}\right]$$

where:

$$\delta = 1 - \phi_1 \rho(1) - \phi_2 \rho(2) - \cdots - \phi_p \rho(p)$$

and

$$\underset{p\times p}{R} = \begin{bmatrix} 1 & \rho(1) & \rho(2) & \cdots & \cdots & \rho(p-1) \\ \rho(1) & 1 & \rho(1) & \rho(2) & \cdots & \rho(p-2) \\ \rho(2) & \rho(1) & 1 & \rho(1) & \cdots & \rho(p-3) \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho(p-2) & \cdots & \cdots & \ddots & 1 & \rho(1) \\ \rho(p-1) & \cdots & \cdots & \cdots & \rho(1) & 1 \end{bmatrix}.$$

Proof. Given that Y_t is stationary we expect a law of large numbers to hold so that:

$$plim \frac{1}{T} \sum_{t=1}^{T} X_t X_t^T = E\left[X_t X_t^T\right].$$

Now you can verify by taking expectations of each of the components of $X_t X_t^T$ that:

$$E\left[X_t X_t^T\right] = \gamma(0) R.$$

But by Theorem 92

$$\begin{aligned}\gamma(0) &= \frac{\sigma^2}{1-\phi_1\rho(1)-\phi_2\rho(2)-\cdots-\phi_p\rho(p)} \\ &= \frac{\sigma^2}{\delta}.\end{aligned}$$

Therefore:

$$\begin{aligned}\sigma^2(\ plim\frac{1}{T}\sum_{t=1}^{T}X_tX_t^T)^{-1} &= \frac{\sigma^2}{\gamma(0)}R^{-1} \\ &= \delta R^{-1}.\end{aligned}$$

■

Remark 106 *Note that R is positive-definite (and hence R^{-1} exists) by Theorem 37.*

3.8.2 Calculating the Log-Likelihood

If Y_t follows an AR(p) process then as we have seen the $Y_t\ 's$ will be correlated across time. This means that the usual method of calculating the likelihood as the product of the likelihoods of each individual observation will not work; in other words:

$$p(Y_T,Y_{T-1},Y_{T-2},\ldots Y_1)\neq p(Y_T)\,p(Y_{T-1})\,p(Y_{T-2})\cdots p(Y_1)\,.$$

Instead we will factor the likelihood by breaking it down into the product of a succession of conditional distributions. This trick is of fundamental importance in time series analysis where the data are not independent across time.

Given that we can write an AR(p) as:

$$Y_t = X_t^T\phi + a_t \tag{3.177}$$

where X_t is given in (3.172), and since:

$$a_t \sim N[0,\sigma^2] \tag{3.178}$$

it follows that if I_{t-1} is the information set at time $t-1$ then:

$$Y_t|I_{t-1} \sim N[X_t^T\phi,\ \sigma^2] \tag{3.179}$$

and so the conditional density is:

$$p(Y_t|I_{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(Y_t-X_t^T\phi)^2}{\sigma^2}\right). \tag{3.180}$$

Given a set of data of T observations of Y_t and X_t for $t = 1, 2, \ldots T$ or equivalently given

$$I_T = \{Y_T, Y_{T-1}, Y_{T-2}, \ldots Y_1, Y_0, Y_{-1}, \ldots Y_{-p+1}\} \tag{3.181}$$

we wish to calculate the density:

$$p\left(Y_T, Y_{T-1}, Y_{T-2}, \ldots Y_1 | \overbrace{Y_0, Y_{-1}, \ldots Y_{-p+1}}^{\text{starting values}}\right) \equiv p(I_T) \tag{3.182}$$

As already noted, since the Y_t $'s$ are not independent we cannot factor the joint density in the usual way. Instead we factor using conditional distribution; i.e. that

$$p\left(A, B\right) = P\left(A|B\right) P\left(B\right) \tag{3.183}$$

with $A = Y_T$ and $B = \{Y_{T-1}, Y_{T-2}, \ldots Y_{-p+1}\} \equiv I_{T-1}$. We thus have:

$$p(I_T) = p(Y_T|I_{T-1})p(I_{T-1}) \tag{3.184}$$

or

$$p(I_T) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(Y_T - X_T^T\phi)}{\sigma^2}\right) p(I_{T-1}). \tag{3.185}$$

We then can do the same thing with $p(I_{T-1})$, i.e.,

$$p(I_{T-1}) = p(Y_{T-1}|I_{T-2})p(I_{T-2}).$$

Continuing this we find that:

$$p(I_T) = \left(2\pi\sigma^2\right)^{-\frac{T}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(Y_t - X_t^T\phi)^2\right) p\left(I_0\right). \tag{3.186}$$

Ignoring the constant term involving $(2\pi)^{-\frac{T}{2}}$ and the asymptotically negligible: $p\left(I_0\right)$, we obtain the (approximate) log-likelihood:

$$\begin{aligned} l\left(\phi, \sigma^2 | I_t\right) &= \ln\left(p(I_T)\right) \qquad (3.187) \\ &= -\frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(Y_t - X_t^T\phi)^2 \end{aligned}$$

or:

Theorem 107 *For an AR(p) model the log-likelihood is given by:*

$$l\left(\phi, \sigma^2\right) = -\frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T} a_t\left[\phi\right]^2 \tag{3.188}$$

where $a_t[\phi]$, the residual for a given ϕ, is defined by:

$$\begin{aligned} a_t[\phi] &\equiv Y_t - X_t^T\phi \qquad (3.189) \\ &\equiv Y_t - \sum_{j=1}^{p}\phi_j Y_{t-j}. \end{aligned}$$

Since maximizing $l\left(\phi,\sigma^2\right)$ with respect to ϕ is identical to minimizing the sum of squares function $\sum_{t=1}^{T} a_t[\phi]^2$, we conclude immediately that:

Theorem 108 *For the AR(p) model the ML estimator and OLS estimators of ϕ are identical or:*

$$\hat{\phi}_{ML} = \hat{\phi} = \left(\sum_{t=1}^{T} X_t X_t^T\right)^{-1} \sum_{t=1}^{T} X_t Y_t.$$

Define the least squares residual $\hat{a}_t$ by:

$$\hat{a}_t \equiv a_t\left[\hat{\phi}\right] = Y_t - X_t^T\hat{\phi} = Y_t - \sum_{j=1}^{p}\hat{\phi}_j Y_{t-j}. \qquad (3.190)$$

Then solving for the maximum likelihood estimator of $\hat{\sigma}^2$ from the first-order conditions:

$$\frac{\partial l\left(\hat{\phi},\hat{\sigma}^2\right)}{\partial\sigma^2} = -\frac{T}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2\left(\hat{\sigma}^2\right)^2}\sum_{t=1}^{\infty}\underbrace{\left(Y_t - \sum_{j=1}^{p}\hat{\phi}_j Y_{t-j}^2\right)^2}_{=\hat{a}_t^2} = 0 \qquad (3.191)$$

we obtain:

Theorem 109 *For the AR(p) model the ML estimator of σ^2 is given by:*

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{a}_t^2.$$

Thus the ML estimate of $\hat{\sigma}^2$ is equal to the sum of squared residuals divided by the number of observations.

If we define $l^* \equiv l\left(\hat{\phi},\hat{\sigma}^2\right)$ as the maximized log-likelihood then:

Theorem 110 *For the AR(p) model the maximized log-likelihood l^* is given by:*

$$l^* = -\frac{T}{2}\ln\left(\hat{\sigma}^2\right) - \frac{T}{2}.$$

Proof. This follows from Theorem 109 since:

$$\begin{aligned} l^* &\equiv l\left(\hat{\phi}, \hat{\sigma}^2\right) = -\frac{T}{2}\ln\left(\hat{\sigma}^2\right) - \frac{1}{2\hat{\sigma}^2}\underbrace{\sum_{t=1}^{T}\hat{a}_t^2}_{=T\hat{\sigma}^2} \\ &= -\frac{T}{2}\ln\left(\hat{\sigma}^2\right) - \frac{T}{2}. \end{aligned}$$

■

3.8.3 Likelihood Ratio Tests

Theorem 110 can be used to derive the likelihood ratio statistic Λ for any set of hypotheses.

Suppose we wish to test a restricted version of an AR process against the alternative of some unrestricted version.

We proceed by estimating the restricted version of the model and obtaining the maximized log-likelihood l_R^* given by:

$$l_R^* = -\frac{T}{2}\ln\left(\hat{\sigma}_R^2\right) - \frac{T}{2} \tag{3.192}$$

where $\hat{\sigma}_R^2$ is the estimator of σ^2 from the restricted model.

We then estimate the unrestricted version of the model and obtain l_U^* given by:

$$l_U^* = -\frac{T}{2}\ln\left(\hat{\sigma}_U^2\right) - \frac{T}{2} \tag{3.193}$$

where $\hat{\sigma}_U^2$ is the estimator of σ^2 from the unrestricted model. We then have:

Theorem 111 *For an AR(p) process the likelihood ratio statistic is:*

$$\begin{aligned} \Lambda &= -2\left(l_R^* - l_U^*\right) \qquad (3.194) \\ &= T\ln\left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2}\right). \end{aligned}$$

As well we have:

Theorem 112 *Under the null that the restrictions are correct:*

$$\Lambda \overset{a}{\sim} \chi_r^2 \tag{3.195}$$

where r is the number of restrictions or the difference in the number of parameters in the restricted and unrestricted models.

An Example

Suppose we believe that Y_t in an AR(2) but wish to apply a diagnostic test to see whether this belief is consistent with the data. To do this we overfit by estimating an AR(5):

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} + \phi_5 Y_{t-5} + a_t \tag{3.196}$$

If our belief of an AR(2) is correct then

$$\phi_3 = \phi_4 = \phi_5 = 0.$$

We therefore test:

$$\begin{aligned} H_0 &: Y_t \sim \text{AR}(2) \quad (\text{or } \phi_3 = \phi_4 = \phi_5 = 0) \ \text{ versus} \\ H_1 &: Y_t \sim \text{AR}(5) \ (\text{or } \phi_3 \neq 0 \text{ or } \phi_4 \neq 0 \text{ or } \phi_5 \neq 0). \end{aligned}$$

Thus the null hypothesis is that the chosen AR(2) model is correct while the alternative implies that an AR(2) is false. This form of overfitting then is a way of deciding if a given model is consistent with the data or not.

If $\sigma_R^2 = \hat{\sigma}_2^2$ is the restricted estimator of σ^2 for the AR(2) model and $\sigma_U^2 = \hat{\sigma}_5^2$ is the unrestricted estimator of σ^2 for the AR(5) model then

$$\Lambda = T \ln\left(\frac{\hat{\sigma}_2^2}{\hat{\sigma}_5^2}\right) \overset{a}{\sim} \chi_3^2. \tag{3.197}$$

3.8.4 Estimating p

We can also use l^* in Theorem 110 to obtain a consistent estimator of p, the order of the autoregressive process. One of earliest estimators is based on the Akaike Information Criterion or *AIC* . The *AIC* is constructed to choose that model which maximizes the trade-off between fit and parsimony given by:

$$AIC \propto \underbrace{\text{maximized log-likelihood}}_{\Rightarrow \text{fit}} \text{ minus } \underbrace{\#\text{ parameters}}_{\Longrightarrow \text{ parsimony}}.$$

If for an AR(k) process the maximized log-likelihood is l_k^* , then the fit term is given by:

$$l_k^* = -\frac{T}{2}\ln\left(\hat{\sigma}_k^2\right) - \frac{T}{2}$$

so that for an AR(k) the Akaike Information Criterion $AIC\,(k)$ is given as:

$$-\frac{T}{2}\ln\left(\hat{\sigma}_k^2\right) - \frac{T}{2} + k. \tag{3.198}$$

Actually this is not the usual way the *AIC* is expressed. We can ignore the constant term $-\frac{T}{2}$ in the maximized log-likelihood (which won't affect anything anyway) and multiply (3.198) by $-\frac{2}{T}$ to obtain:

Definition 113 *For an AR(k) process the Akaike Information Criterion is:*

$$AIC(k) = \ln\left(\hat{\sigma}_k^2\right) + \frac{2k}{T}.$$

Note that since we have multiplied by a negative number, $AIC(k)$ needs to be *minimized* now.

We have:

Definition 114 *The AIC estimator of p is that value of k which minimizes $AIC(k)$ or*

$$\hat{p}_{AIC} = \arg\min_k AIC(k) \qquad k = 0, 1, 2, \ldots p_{\max}\ . \tag{3.199}$$

It can be shown that as $T \to \infty$ that in probability:

$$\hat{p}_{AIC}\ \to \tilde{p} > p \tag{3.200}$$

so that AIC tends to overestimate the true p. This theoretical result corresponds to the fact that in practice the AIC often seems to pick unreasonably large values of p.

This problem can be remedied by giving the parsimony term involving the number of parameters a greater weight. One way to do this is to replace the 2 in $\frac{2k}{T}$ with $\ln(T)$ to obtain the Schwarz criterion[3] or $SC\left(k\right)$. We then have:

Definition 115 *For an AR(k) process the Schwarz criterion is:*

$$SC(k) = \ln\left(\hat{\sigma}_k^2\right) + \frac{\ln(T)k}{T}.$$

We estimate p as using:

Definition 116 *The Schwarz estimator of p is that value of k which minimizes $SC(k)$ or:*

$$\hat{p}_{SC} \equiv \arg\min_k SC(k)\ ,\ \ k = 0, 1, 2, \ldots p_{\max}. \tag{3.201}$$

It can be shown that in probability as $T \to \infty$

$$\hat{p}_{SC}\ \to p \tag{3.202}$$

so the Schwarz criterion selects the correct p asymptotically.

Remark 117 *A common error when using either $AIC\left(k\right)$ or $SC\left(k\right)$ is to forget to include $k = 0$ This corresponds to the case where $Y_t = a_t$ or white noise so that*

$$\hat{\sigma}_0^2 = \frac{1}{T}\sum_{t=1}^{T} Y_t^2 \tag{3.203}$$

and

$$AIC(0) = SC(0) = \ln(\hat{\sigma}_0^2). \tag{3.204}$$

[3] This is sometimes refered to as the Bayesian Information Criterion.

3.9 The Partial Autocorrelation Function

For an AR(p) process the autocorrelation function $\rho(k)$ is a damped exponential. Although $\rho(k)$ converges to zero, it never actually reaches or becomes zero. The partial autocorrelation ϕ_{kk}, on the other hand function, does actually reach zero, and it does so precisely after $k = p$. In particular we have:

Theorem 118 *If Y_t is an AR(p) then*

$$\begin{aligned} \phi_{kk} &\neq 0, \text{ for } k = 1, 2, \dots p \\ \phi_{kk} &= 0, \text{ for } k = p+1, p+2, \dots \infty. \end{aligned}$$

Proof. For an AR(p) process the optimal one-step ahead forecast using the entire past history of Y_t is given by:

$$E_t[Y_{t+1}] = \sum_{j=0}^{p-1} \phi_j Y_{t-j}. \tag{3.205}$$

Note this is a function only of the last p values of Y_t, Y_{t-1}, $Y_{t-2} \dots Y_{t-p+1}$. Suppose now that we use the last k values of Y_t to forecast Y_{t+1} and that $k > p$. Then the optimal weights can be calculated from (2.53) or the Yule-Walker equations in Theorem 64 as:

$$\begin{aligned} E[Y_t | Y_t, Y_{t-1}, Y_{t-2}, \dots Y_{t-k+1}] &= \sum_{j=0}^{k-1} \phi_{kj} Y_{t-j} \qquad (3.206) \\ &= \sum_{j=0}^{p-1} \phi_j Y_{t-j}. \end{aligned}$$

Combining (3.205) and (3.206)it follows that:

$$\begin{aligned} \phi_{jk} &= \phi_j, \text{ for } j = 1, 2, \dots p \qquad (3.207) \\ \phi_{jk} &= 0, \text{ for } j = p+1, p+2, \dots k. \end{aligned}$$

Now by Definition 68 the partial autocorrelation is ϕ_{kk} so we conclude that $\phi_{kk} = 0$ for $k > p$. ■

For example suppose Y_t is an AR(2) then we know that the optimal forecast of Y_{t+1} is given by:

$$E_t[Y_{t+1}] = \phi_1 Y_t + \phi_2 Y_{t-1}. \tag{3.208}$$

If we were then to calculate:

$$E[Y_{t+1} | Y_t, Y_{t-1}, Y_{t-2}, Y_{t-3}] = \phi_{41} Y_t + \phi_{42} Y_{t-1} + \phi_{43} Y_{t-2} + \phi_{44} Y_{t-3} \tag{3.209}$$

since this forecast can be no better than the optimal, it must be that:

$$\begin{aligned} \phi_{41} &= \phi_1, \ \phi_{42} = \phi_2 \qquad (3.210) \\ \phi_{43} &= 0, \ \phi_{44} = 0. \end{aligned}$$

In particular $\phi_{44} = 0$.

In general for an AR(2) it must be that:

$$0 = \phi_{33} = \phi_{44} = \phi_{55} = \cdots . \tag{3.211}$$

Chapter 4

ARMA(p,q) Models

AR(p)$'s$ form a very general class of stochastic processes that is nearly sufficient for applied work. Nevertheless it turns out that broadening this class of time series models to include moving average or MA(q)'s, or mixed ARMA(p,q)'s is very useful. Sometimes, for example, we can model a particular time series as an MA(q) with fewer parameters than if we modelled it as an AR(p). Other times theory predicts other models besides AR(p)'s. For example rational expectations predicts that forecast errors will be moving average processes while measurement error or aggregation often mean that even if the underlying process is an AR(p) that the observed process will be an ARMA(p,q). Thus if Y_t is an AR(p) process but we observe $Y_t^* = Y_t + e_t$ where e_t is a white noise measurement error, then Y_t^* is an ARMA(p,p) process.

We first consider the class of moving average processes or MA(q)'s.

4.1 MA(q) Processes

4.1.1 Introduction

As we have already seen, stationary AR(p)$'s$ are characterized by short-memory exponential decay with ψ_k, $\rho(k)$ and $E_t[Y_{t+k}]$ all decaying exponentially to 0 as $k \to \infty$. Exponential decay means that these functions never quite reach zero, no matter how large is k, in the same manner that if you eat half of the pie each day there will always be a little pie left.

In contrast moving average processes or MA(q)'s are characterized by a finite cutoff; that is ψ_k, $\rho(k)$ and $E_t[Y_{t+k}]$ actually reach 0 and then stay at 0. Furthermore this cutoff occurs when $k = q + 1$.

We define moving averages as follows:

Definition 119 *We say that Y_t ~MA(q) or Y_t follows a q^{th} order moving average process if:*

$$Y_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q},$$

or alternatively:

$$Y_t = \theta\left(B\right) a_t$$

where:

$$\begin{aligned}\theta(B) &= \theta_o + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \\ &= \sum_{j=0}^{q} \theta_j B^j,\end{aligned}$$

and where: $\theta_o \equiv 1$.

Example 120 *The process:*

$$Y_t = a_t + \theta_1 a_{t-1} \tag{4.1}$$

is an MA(1) with

$$\theta\left(B\right) = 1 + \theta_1 B$$

while:

$$Y_t = a_t + \theta_1 a_{t-2} + \theta_2 a_{t-2} \tag{4.2}$$

is an MA(2) with

$$\theta\left(B\right) = 1 + \theta_1 B + \theta_2 B.$$

4.1.2 Stationarity

The first question we might ask is under what circumstances are MA(q)'s stationary? While you might think that it has something to do with the roots of $\theta\left(B\right)$, in fact the correct answer is much simpler. We have:

Theorem 121 *MA(q)'s are* always *stationary with:*

$$\gamma(0) \equiv Var[Y_t] = \sigma^2\left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right) < \infty.$$

The Wold representation for an MA(q) is also easy to derive. Since the Wold representation takes the form:

$$Y_t = \psi\left(B\right) a_t \tag{4.3}$$

and since for an MA(q):

$$Y_t = \theta\left(B\right) a_t; \tag{4.4}$$

it follows (trivially) that $\psi\left(B\right) = \theta\left(B\right)$ or

Theorem 122 *For an MA(q):*

$$\psi_k = \begin{cases} \theta_k \; for \;\; k = 0, 1, 2, \ldots q \\ \quad 0 \quad for \; k > q. \end{cases} \tag{4.5}$$

Remark 123 *Note that ψ_k has the cutoff property which is typical of MA(q)'s. This is unlike $AR(p)'s$, where the ψ_k 's decay exponentially to zero but never actually reach zero.*

Example 124 *For example the MA(1)*:

$$Y_t = a_t + 0.5a_{t-1} \tag{4.6}$$

has:

$$\psi_0 = 1, \psi_1 = 0.5, \;\; and \; \psi_k = 0 \; for \; k = 2, 3, \ldots \infty.$$

Since ψ_k actually reaches 0 at $k = q + 1$ and stays there forever, it is trivial that:

Theorem 125 *ψ_k has the short-memory property or $\psi_k = O\left(\tau^k\right)$.*

4.1.3 The Autocorrelation Function

The autocorrelation function is also characterized by a finite cutoff. We have:

Theorem 126 *If Y_t ∼MA(q):*

$$\rho(k) = \begin{cases} \frac{\sum_{j=0}^{q-|k|} \theta_j \theta_{j+k}}{\sum_{j=0}^{q} \theta_j^2} & k = 0, 1, 2, \ldots q \\ = 0 & k > q. \end{cases}$$

Proof. This follows from (2.40) and the fact that $\psi_k = \theta_k$. ■

Again since $\rho(k)$ actually reaches 0 at $k = q + 1$ and stays there forever, it is trivial that:

Theorem 127 *$\rho(k)$ has the short-memory property or $\rho(k) = O\left(\tau^k\right)$.*

4.1.4 Forecasting

This finite cutoff property also holds for forecasts. Thus:

Theorem 128 *If Y_t ∼MA(q) then*

$$E_t[Y_{t+k}] = \begin{cases} \sum_{j=0}^{q-k} \theta_{k+j} a_{t-j} \;\; for \;\; k = 0, 1, 2, \ldots q \\ = 0 \; for \quad k > q. \end{cases} \tag{4.7}$$

Again since $E_t[Y_{t+k}]$ actually reaches 0 at $k = q + 1$ and stays there forever, it is trivial that:

Theorem 129 *$E_t[Y_{t+k}]$ has the short-memory property or $E_t[Y_{t+k}] = O\left(\tau^k\right)$.*

We can then find $Var_t\left[Y_{t+k}\right]$ as:

Theorem 130 *If $Y_t \sim$MA(q) then:*

$$\begin{aligned} Var_t\left[Y_{t+k}\right] &= \sigma^2\left(1+\theta_1^2+\cdots+\theta_{k-1}^2\right) \text{ for } k=1,2,\ldots q \\ &= \gamma(0) \text{ for } k>q. \end{aligned}$$

4.1.5 An MA(1) Example

Consider the MA(1):

$$Y_t = a_t + 0.5a_{t-1}, \sigma = 0.05. \tag{4.8}$$

Then:

$$\psi_1 = 0.5 \text{ and } \psi_k = 0 \text{ for } k \geq 2.$$

We can calculate the variance $\gamma(0)$ as:

$$\begin{aligned} \gamma(0) &= \sigma^2\left(1+\theta_1^2\right) && (4.9) \\ &= 0.05^2\left(1+0.5^2\right) && (4.10) \\ &= 0.003125 \end{aligned}$$

while:

$$\begin{aligned} \rho(1) &= \frac{\theta_1}{1+\theta_1^2} && (4.11) \\ &= \frac{0.5}{1+0.5^2} \\ &= 0.4 \end{aligned}$$

with $\rho(k) = 0$ for $k = 2, 3, \ldots \infty$.

If $a_t = 0.04$ then the one-step ahead forecast is:

$$\begin{aligned} E_t\left[Y_{t+1}\right] &= \theta_1 a_t && (4.12) \\ &= 0.5 \times 0.04 \\ &= 0.02 \end{aligned}$$

while for horizons greater than one the forecast is the unconditional mean of 0 or:

$$E_t\left[Y_{t+k}\right] = 0 \text{ for } k = 2, 3, \ldots. \tag{4.13}$$

To construct confidence intervals for the forecasts we need:

$$\begin{aligned} Var_t\left[Y_{t+1}\right] &= \sigma^2 = 0.05^2 && (4.14) \\ Var_t\left[Y_{t+k}\right] &= \gamma(0) = \sigma^2\left(1+\theta_1^2\right) = 0.003125 \text{ for } k \geq 2 \end{aligned}$$

and so confidence intervals for our forecasts would be:

Forecast Confidence Intervals

$k=1$	$0.02 \pm 1.96(0.05) = 0.02\pm\ .098$
$k=2$	$0 \pm 1.96\,(0.003125)^{1/2} = 0\pm\ 0.11$
$k=3$	$0 \pm 1.96\,(0.003125)^{1/2} = 0\pm\ 0.11$
$k=4$	$0 \pm 1.96\,(0.003125)^{1/2} = 0\pm\ 0.11$

.

4.1.6 An MA(2) Example

Now consider the MA(2):

$$Y_t = a_t + 0.5a_{t-1} + 0.4a_{t-1},\ \sigma = 0.05 \tag{4.15}$$

The infinite moving average weights are then:

$$\begin{aligned} \psi_1 &= 0.5, && (4.16)\\ \psi_2 &= 0.4, && (4.17)\\ \psi_k &= 0,\ \text{for } k = 3, 4, \ldots \infty. \end{aligned}$$

Thus:

$$\begin{aligned} \gamma\,(0) &= \sigma^2\left(1+\theta_1^2+\theta_2^2\right) && (4.18)\\ &= 0.05^2\left(1+0.5^2+0.4^2\right)\\ &= 0.003525 \end{aligned}$$

while:

$$\begin{aligned} \rho\,(1) &= \frac{\theta_0\theta_1+\theta_1\theta_2}{1+\theta_1^2+\theta_2^2} && (4.19)\\ &= \frac{0.5+0.5\times 0.4}{1+0.5^2+0.4^2}\\ &= 0.496 \end{aligned}$$

and

$$\begin{aligned} \rho\,(2) &= \frac{\theta_0\theta_2}{1+\theta_1^2+\theta_2^2} && (4.20)\\ &= \frac{0.4}{1+0.5^2+0.4^2}\\ &= 0.283 \end{aligned}$$

with

$$0 = \rho\,(3) = \rho\,(4) = \cdots\ . \tag{4.21}$$

If $a_t = 0.04$ and $a_{t-1} = 0.02$ then:

$$\begin{aligned} E_t\left[Y_{t+1}\right] &= \theta_1 a_t + \theta_2 a_{t-1} \\ &= 0.5 \times 0.04 + 0.4 \times 0.02 \\ &= 0.028 \end{aligned} \tag{4.22}$$

while:

$$\begin{aligned} E_t\left[Y_{t+2}\right] &= \theta_2 a_t \\ &= 0.4 \times 0.04 \\ &= 0.016 \end{aligned} \tag{4.23}$$

with $E_t\left[Y_{t+k}\right] = 0$ for $k = 3, 4, \ldots$. To construct confidence intervals for our forecasts we need:

$$\begin{aligned} Var_t\left[Y_{t+1}\right] &= \sigma^2 = 0.05^2 \\ Var_t\left[Y_{t+2}\right] &= \sigma^2\left(1+\theta_1^2\right) = 0.003125 \\ Var_t\left[Y_{t+k}\right] &= \gamma\left(0\right) = 0.003525 \text{ for } k \geq 3 \end{aligned} \tag{4.24}$$

and so the confidence intervals would be:

Forecast Confidence Intervals

$k = 1$	$0.028 \pm 1.96(0.05) = 0.028 \pm 0.098$
$k = 2$	$0.016 \pm 1.96\left(0.003125\right)^{1/2} = 0.016 \pm 0.11$
$k = 3$	$0 \pm 1.96\left(0.003525\right)^{1/2} = 0 \pm 0.116$
$k = 4$	$0 \pm 1.96\left(0.003525\right)^{1/2} = 0 \pm 0.116$

.

4.2 Invertibility

4.2.1 Definition

By the Wold representation a stationary time series can be represented as an MA(∞) with a finite variance, which is why an MA(q) is always stationary. An AR(p) is not always stationary because it is sometimes not legitimate to throw $\phi\left(B\right)$ on to the right-hand side as

$$\phi\left(B\right) Y_t = a_t \Longrightarrow Y_t = \frac{1}{\phi\left(B\right)} a_t \tag{4.25}$$

if the roots of $\phi\left(B\right)$ are not all greater than 1 in absolute value.

One might also ask the question if a time series has an AR(∞) representation. We define series which have an AR(∞) representation as invertible time series as:

Definition 131 ***Invertibility:*** *A process is invertible if it has an infinite autoregressive representation.*

$$\pi\left(B\right) Y_t = a_t$$

where:

$$\pi(B) = 1 + \pi_1 B + \pi_2 B^2 + \cdots$$

and where $\pi_k \to 0$ as $k \to \infty$.

4.2.2 Conditions for Invertibility

For an invertible MA(q) can we throw $\theta(B)$ on to the left-hand side as

$$Y_t = \theta(B)\, a_t \Longrightarrow \frac{1}{\theta(B)} Y_t = a_t \tag{4.26}$$

so that:

$$\pi(B) = \frac{1}{\theta(B)}.$$

Just as with stationarity this turns out to be legitimate only if $\theta(B)$ has all its roots greater than 1 in absolute value or

Theorem 132 *An MA(q) is invertible only if*

$$\theta(B) = 0 \Longrightarrow |B| > 1$$

in which case $\pi_k = O\left(\tau^k\right)$ has the short-memory property where:

$$\pi_k = \sum_{j=1}^{q} D_j r_j^k,$$

$\theta\left(r_j^{-1}\right) = 0$ *and* $\tau = \max_j |r_j| < 1$.

Again, just as MA(q)'s are always stationary we have:

Theorem 133 *An AR(p) process $\phi(B)\, Y_t = a_t$ is always invertible with:*

$$\begin{aligned} \pi_k &= -\phi_k, \text{ for } k = 0, 1, \ldots p \\ &= 0, \text{ for } k > p. \end{aligned}$$

Example 134 *An example of an invertible MA(1) process is:*

$$Y_t = a_t - 0.5 a_{t-1} \tag{4.27}$$

since it has an infinite AR representation:

$$\frac{1}{1 - 0.5B} Y_t = a_t \tag{4.28}$$

or:

$$Y_t + 0.5 Y_{t-1} + 0.5^2 Y_{t-2} + 0.5^3 Y_{t-3} + \cdots = a_t \tag{4.29}$$

where:

$$\pi_k = (0.5)^k. \tag{4.30}$$

Example 135 *An example of a non-invertible MA(1) process is:*

$$Y_t = a_t - 2a_{t-1} \tag{4.31}$$

since it would have an infinite AR representation given by

$$\frac{1}{1-2B} Y_t = a_t \tag{4.32}$$

or:

$$Y_t + 2Y_{t-1} + 2^2 Y_{t-2} + 2^3 Y_{t-3} + \cdots = a_t \tag{4.33}$$

where:

$$\pi_k = (2)^k \tag{4.34}$$

does not converge to zero as $k \to \infty$.

Invertibility is not as important as stationarity, but it does arise in estimation as many methods of approximating the log-likelihood require invertibility.

4.2.3 Why MA(q)'s are Almost Always Invertible

A curious fact about MA(q) processes is that although:

$$Y_t = a_t + 3a_{t-1} \tag{4.35}$$

is not invertible, there is an observationally equivalent representation:

$$Y_t = \tilde{a}_t + \frac{1}{3}\tilde{a}_{t-1} \tag{4.36}$$

which is invertible where

$$Var\left[\tilde{a}_t\right] = 3^2 Var\left[a_t\right]. \tag{4.37}$$

More generally we have

Theorem 136 *If:*

$$Y_t = a_t + \theta a_{t-1}$$

with $Var\left[a_t\right] = \sigma^2$ then an observationally equivalent MA(1) is:

$$Y_t = \tilde{a}_t + \tilde{\theta}\tilde{a}_{t-1}$$

where: $Var\left[\tilde{a}_t\right] = \theta^2\sigma^2$ and $\tilde{\theta} = \frac{1}{\theta}$. You can verify the truth of this theorem by showing the two $MA(1)'s$ have identical autocovariance functions.

This result means that if $\theta > 1$ for the first representation, and hence the MA(1) is not invertible, then there exists a second MA(1) with a MA coefficient $\frac{1}{\theta} < 1$ which is invertible.

Analogous results hold for higher order MA(q)'s, in particular:

Proposition 137 *Given an MA(q) with:* $Y_t = \theta(B)\, a_t$ *where* $\theta(B)$ *has all real roots and can be factored as:*

$$\theta(B) = (1 - r_1 B)(1 - r_2 B) \cdots (1 - r_q B)$$

then there is an observationally equivalent representation: $Y_t = \tilde{\theta}(B)\, \tilde{a}_t$ *where:*

$$\tilde{\theta}(B) = \left(1 - (r_1)^{\pm 1} B\right)\left(1 - (r_2)^{\pm 1} B\right) \cdots \left(1 - (r_q)^{\pm 1} B\right).$$

Thus if $\theta(B)$ is non-invertible because $|r_i| > 1$ we can replace r_i by r_i^{-1} in $\tilde{\theta}(B)$ and obtain an invertible MA(q). The same argument applies to complex roots except one must treat roots then in conjugate pairs.

Because of this curious property of MA(q)'s we can almost always avoid the invertibility problem by picking the invertible representation.

Problems can still arise, however, when $|r_i| = 1$ since in this case $\left|r_i^{-1}\right| = 1$ as well. In particular for the MA(1) if $\theta = 1$ or $\theta = -1$ then there is no way around non-invertibility.

This sometimes occurs in practice, especially if we over-difference. For example suppose the truth is TS with a white noise cycle or $Y_t = a_t$, so that:

$$X_t = \alpha + \mu t + a_t \tag{4.38}$$

but we wrongly assume that the process is DS. In this case we would arrive at a cycle:

$$\begin{aligned} Y_t &= X_t - X_{t-1} - \mu \\ &= a_t - a_{t-1} \end{aligned} \tag{4.39}$$

which is a non-invertible MA(1) with $\theta = -1$.

We can summarize the relationship between stationarity and invertibility for AR(p) and MA(q) models as follows:

Duality Between Stationarity and Invertibility

Model	Stationarity	Invertiblity
AR(p)	$=$MA(∞)	Always
Condition:	Roots of $\phi(B)^*$	None
MA(q)	Always	$=$AR(∞)
Condition:	None	Roots of $\theta(B)^*$

*The roots must be greater than 1 in absolute value.

4.3 Mixed Processes

4.3.1 Introduction

So far we have considered pure AR(p) and pure MA(q) processes. We can combine the two to form a broader class of models where there is both an MA and an AR component. This is called an ARMA(p,q) or autoregressive moving average which we define as:

Definition 138 *We say that Y_t ∼ARMA(p,q) if*

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + a_t + \sum_{j=1}^{q} \theta_j a_{t-j}. \tag{4.40}$$

or equivalently

$$\phi\left(B\right) Y_t = \theta\left(B\right) a_t \tag{4.41}$$

and neither $\phi\left(B\right)$ nor $\theta\left(B\right)$ share any common roots.

Remark 139 *The condition that $\phi\left(B\right)$ and $\theta\left(B\right)$ not share any common roots is required for identification. If $\phi\left(B\right)$ and $\theta\left(B\right)$ shared a common root, say $\tilde{r}$, then we could write*

$$\begin{aligned}\phi\left(B\right) &= \left(1-\tilde{r}B\right)\tilde{\phi}\left(B\right)\\ \theta\left(B\right) &= \left(1-\tilde{r}B\right)\tilde{\theta}\left(B\right)\end{aligned}$$

so that the ARMA(p,q) model becomes:

$$\left(1-\tilde{r}B\right)\tilde{\phi}\left(B\right)Y_t = \left(1-\tilde{r}B\right)\tilde{\theta}\left(B\right)a_t.$$

Cancelling $\left(1-\tilde{r}B\right)$ from both sides then leaves:

$$\tilde{\phi}\left(B\right)Y_t = \tilde{\theta}\left(B\right)a_t$$

which is an observationally equivalent ARMA(p-1,q-1) model. Furthermore given any ARMA(p,q):

$$\phi\left(B\right)Y_t = \theta\left(B\right)a_t$$

then we can always multiply both sides by $\left(1-\tilde{r}B\right)$ to obtain:

$$\left(1-\tilde{r}B\right)\phi\left(B\right)Y_t = \left(1-\tilde{r}B\right)\theta\left(B\right)a_t$$

for an arbitrary $\tilde{r}$ to obtain an observationally equivalent ARMA(p+1,q+1) process.

4.3.2 The Wold Representation

To calculate the Wold or infinite moving average representation simply throw $\phi(B)$ on the right-hand side of (4.41) to obtain:

$$Y_t = \frac{\theta(B)}{\phi(B)} a_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots . \tag{4.42}$$

To calculate ψ_k recursively we use the following:

Theorem 140 *If Y_t ∼ARMA(p,q) the ψ_k 's can be calculated recursively as:*

$$\begin{aligned} \psi_k &= \sum_{j=1}^{p} \phi_j \psi_{k-j} + \theta_k \quad , \quad k = 0, 1, 2, \dots q \qquad (4.43) \\ &= \sum_{j=1}^{p} \phi_j \psi_{k-j} \ , \quad k > q \end{aligned}$$

with starting values $\psi_0 = 1$ and $\psi_k = 0$ for $k < 0$.

Since for $k > q$ we have:

$$\psi_k = \sum_{j=1}^{p} \phi_j \psi_{k-j}$$

it follows that ψ_k follows a linear p^{th} order difference equation. This difference equation is identical to that which we derived for the AR(p) and hence will have the same solution. We therefore have:

Theorem 141 *If Y_t ∼ARMA(p,q) with $p > 0$ the ψ_k 's can be for $k > q$ written as:*

$$\psi_k = \sum_{j=1}^{p} A_j r_j^k$$

where:

$$\phi\left(r_j^{-1}\right) = 0 \text{ for } j = 1, 2, \dots p.$$

From this we conclude that the stationarity of an ARMA(p,q) depends only on the roots of $\phi(B)$. In particular:

Theorem 142 *For a stationary ARMA(p,q) process $\psi_k = O\left(\tau^k\right)$ has the short-memory property or:*

$$|\psi_k| \leq A\tau^k$$

where $0 \leq \tau = \max_j [|r_j|] < 1$ and where $\phi(r_j) = 0$.

4.3.3 The Autocorrelation Function

A similar result holds for the autocorrelation function. We have:

Theorem 143 *If $Y_t \sim$ARMA(p,q) with $p > 0$ then for $0 \leq k \leq q$*

$$\begin{aligned}\rho(k) &= \phi_1 \rho(k-1) + \phi_2 \rho(k-2) + \cdots + \phi_p \rho(k-p) \\ &\quad + \frac{\sum_{j=1}^{q} \theta_j \psi_{j-k}}{\sum_{j=0}^{\infty} \psi_j^2}\end{aligned} \tag{4.44}$$

while for $k > q$

$$\rho(k) = \phi_1 \rho(k-1) + \phi_2 \rho(k-2) + \cdots + \phi_p \rho(k-p)$$

with starting values given by $\rho(-k) = \rho(k)$.

Proof. The proof follows that of Theorem 94 and using the fact that

$$\sum_{j=1}^{q} \theta_j E[a_{t-j} Y_{t-k}] = \sigma^2 \sum_{j=1}^{q} \theta_j \psi_{k-j}$$

and from Theorem 56 that

$$\gamma(0) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2$$

and using Theorem 2.40. ■

Since the term in the second line of (4.44) becomes zero when $k > q$ and so reduces to:

$$\rho(k) = \phi_1 \rho(k-1) + \phi_2 \rho(k-2) + \cdots + \phi_p \rho(k-p)$$

it follows that $\rho(k)$ follows the same linear p^{th} order difference equation as ψ_k, except of course that the starting values are different. It follows then that

Theorem 144 *If $Y_t \sim$ARMA(p,q) with $p > 0$ then for $k > q$*

$$\rho(k) = \sum_{j=1}^{p} B_j r_j^k$$

where $\phi\left(r_j^{-1}\right) = 0$ for $j = 1, 2, \ldots p$.

It follows immediately that:

Theorem 145 *For a stationary ARMA(p,q) process $\rho(k) = O\left(\tau^k\right)$ has the short-memory property.*

4.3.4 Forecasting

To calculate forecasts for Y_{t+k} recursively we use apply E_t to both sides of (4.40) to obtain:

Theorem 146 *If Y_t ∼ARMA(p,q) with $p > 0$ then for $k > q$*

$$E_t\left[Y_{t+k}\right] = \sum_{j=1}^{p} \phi_j E_t\left[Y_{t+k-j}\right] + \sum_{j=1}^{q} \theta_j E_t\left[a_{t+k-j}\right] \tag{4.45}$$

where

$$E_t\left[Y_{t+k}\right] = Y_{t+k} \text{ and } E_t\left[a_{t+k}\right] = a_{t+k} \text{ for } k \leq 0.$$

Just as with ψ_k and $\rho(k)$ the final term in (4.45) becomes 0 for $k > q$ and so we have:

$$E_t\left[Y_{t+k}\right] = \sum_{j=1}^{p} \phi_j E_t\left[Y_{t+k-j}\right] \text{ for } k > q.$$

From this it follows that:

Theorem 147 *If Y_t ∼ARMA(p,q) with $p > 0$ then*

$$E_t\left[Y_{t+k}\right] = \sum_{j=1}^{p} C_{jt} r_j^k$$

where C_{jt} is a function of the information set at time t.

It follows immediately that

Theorem 148 *For a stationary ARMA(p,q) process $E_t\left[Y_{t+k}\right] = O\left(\tau^k\right)$ has the short-memory property.*

Confidence intervals for forecasts of an ARMA(p,q) process can be calculated as usual from the Wold representation using:

Theorem 149 *If Y_t ∼ARMA(p,q) with $p > 0$ then for $k > q$*

$$Var_t\left[Y_{t+k}\right] = \sigma^2 \sum_{j=0}^{k-1} \psi_j^2 \tag{4.46}$$

where ψ_j is calculated from (4.43).

As $k \to \infty$ the conditional forecast variance approaches the conditional forecast variance very rapidly or: $Var_t\left[Y_{t+k}\right] \to \gamma(0)$. In particular we can show that the difference between $Var_t\left[Y_{t+k}\right]$ and $\gamma(0)$ has the short memory property.

Theorem 150 *If $Y_t \sim$ARMA(p,q) is stationary then:*

$$\left|Var_t\left[Y_{t+k}\right]-\gamma(0)\right| = O\left(\tau^{2k}\right).$$

Proof. Since $|\psi_k| \leq A\tau^k$ and

$$\gamma(0) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2$$

we have:

$$\begin{aligned}\left|Var_t\left[Y_{t+k}\right]-\gamma(0)\right| &\leq \sigma^2 A^2 \sum_{j=k+1}^{\infty} \tau^{2j} \\ &= \left(\frac{\sigma^2 A^2 \tau^2}{1-\tau^2}\right)\tau^{2k}.\end{aligned}$$

■

It is possible the calculate $Var_t\left[Y_{t+k}\right]$ recursively since from Theorem 149

Theorem 151 *If $Y_t \sim$ARMA(p,q) with $p > 0$ then for $k > q$ then:*

$$Var_t\left[Y_{t+k}\right] = Var_t\left[Y_{t+k-1}\right] + \sigma^2\psi_{k-1}^2$$

with $Var_t\left[Y_{t+1}\right] = \sigma^2$.

We then construct a confidence interval for our forecasts in the usual way as:

Theorem 152 *A 95% confidence interval for Y_{t+k} is:*

$$E_t\left[Y_{t+k}\right] \pm 1.96\sqrt{Var_t\left[Y_{t+k}\right]} \tag{4.47}$$

where $E_t\left[Y_{t+k}\right]$ is calculated from Theorem 146 and $Var_t\left[Y_{t+k}\right]$ from Theorem 149 or Theorem 151.

4.3.5 The ARMA(1,1) process

For an ARMA(1,1) process

$$Y_t = \phi Y_{t-1} + a_t + \theta a_{t-1} \tag{4.48}$$

we have:

$$\psi_1 = \phi\psi_0 + \theta = \phi + \theta$$

and for $k \geq 2$

$$\begin{aligned}\psi_k &= \phi\psi_{k-1} &(4.49)\\ &= \phi^{k-1}\psi_1 &(4.50)\\ &= \phi^{k-1}(\phi+\theta). &(4.51)\end{aligned}$$

Thus ψ_k decays exponentially for $k > 1$.

From this we can calculate $\gamma(0)$ as:

$$\begin{aligned}\gamma(0) &= \sigma^2\left(1+\sum_{k=1}^{\infty}\psi_k^2\right) = \sigma^2\left(1+(\theta+\phi)^2\sum_{k=1}^{\infty}\phi^{2(k-1)}\right) \\ &= \sigma^2\left(1+\frac{(\theta+\phi)^2}{1-\phi^2}\right)\end{aligned} \tag{4.52}$$

and $\gamma(1)$ from:

$$\begin{aligned}\gamma(1) &= E\left[Y_t Y_{t-1}\right] = E\left[\left(\phi Y_{t-1}+a_t+\theta a_{t-1}\right)Y_{t-1}\right] \\ &= \phi\gamma(0)+\theta\sigma^2 \\ &= \sigma^2\left(\phi+\theta+\frac{\phi(\theta+\phi)^2}{1-\phi^2}\right).\end{aligned} \tag{4.53}$$

Thus:

$$\rho(1) = \frac{\left(\phi+\theta+\frac{\phi(\theta+\phi)^2}{1-\phi^2}\right)}{\left(1+\frac{(\theta+\phi)^2}{1-\phi^2}\right)} \tag{4.54}$$

and $\rho(k)$ is:

$$\rho(k) = \phi^{k-1}\rho(1) \quad \text{for } k = 1, 2, \ldots\infty. \tag{4.55}$$

Thus $\rho(k)$ decays exponentially for $k > 1$.

For forecasting we have:

$$E_t\left[Y_{t+k}\right] = \phi E_t\left[Y_{t+k-1}\right] + \theta E_t\left[a_{t+k-1}\right] \tag{4.56}$$

so that:

$$E_t\left[Y_{t+k}\right] = \phi Y_t + \theta a_t \tag{4.57}$$

and for $k \geq 2$

$$E_t\left[Y_{t+k}\right] = \phi E_t\left[Y_{t+k-1}\right] = \phi^{k-1}\left(\phi Y_t + \theta a_t\right). \tag{4.58}$$

To construct confidence intervals for an ARMA(1,1) we have:

$$\begin{aligned}Var_t\left[Y_{t+k}\right] &= \sigma^2\left(1+\sum_{j=1}^{k-1}\psi_j^2\right) \\ &= \sigma^2\left(1+\frac{(\theta+\phi)^2\left(1-\phi^{2(k-1)}\right)}{1-\phi^2}\right).\end{aligned} \tag{4.59}$$

4.3.6 Common AR and MA factors

Now consider an ARMA(1,1) model with $\phi = -\theta$ so that

$$Y_t = \phi Y_{t-1} + a_t - \phi a_{t-1}.$$

For this process when we calculate the Wold representation we find that for $k \geq 1$:

$$\psi_k = \phi^{k-1} (\phi + \theta) = \phi^{k-1} (-\theta + \theta) = 0 \tag{4.60}$$

so that

$$Y_t = a_t; \tag{4.61}$$

that is, this process is not really ARMA(1,1) but white noise or ARMA(0,0).

Another way of seeing this is to write the model with $\theta = -\phi$ as :

$$(1 - \phi B) Y_t = (1 - \phi B) a_t \tag{4.62}$$

so that the two polynomials cancel and we are left with:

$$Y_t = a_t \tag{4.63}$$

Whenever you deal with mixed ARMA(p,q) processes you need to be aware of the potential problem of near or exact cancelation. For example if you estimate the ARMA(1,1) process:

$$\begin{aligned} Y_t &= 0.53 Y_{t-1} + a_t - 0.51 a_{t-1} \text{ or} && (4.64) \\ (1 - 0.53B) Y_t &= (1 - 0.51B) a_t && (4.65) \end{aligned}$$

the AR and MA terms nearly cancel and the data is probably best modelled as white noise. It is very likely that such near cancelation will be associated with numerical problems such a failure to achieve convergence with iterative nonlinear estimation procedures.

It is not always obvious that there is a problem with near cancellations. Consider for example the apparent ARMA(2,1) model:

$$Y_t = 0.3 Y_{t-1} + 0.4 Y_{t-2} + a_t + 0.5 a_{t-1} \tag{4.66}$$

which can be written as:

$$\left(1 - 0.3B - 0.4B^2\right) Y_t = (1 + 0.5B) a_t. \tag{4.67}$$

Factoring the AR polynomial we obtain

$$(1 - 0.8B)(1 + 0.5B) Y_t = (1 + 0.5B) a_t \tag{4.68}$$

so that cancelling $(1 + 0.5B)$ from both sides:

$$(1 - 0.8B) Y_t = a_t$$

and so the process is really an AR(1).

4.3.7 Forecasting Growth Rates

So far we have concentrated our forecasting on Y_t. In the real work however most forecasting exercises are concerned with growth rates and not the TS or DS cycle Y_t. It is important therefore to be able to convert forecasts involving Y_t into forecasts involving growth rates.

Suppose we are interested in forecasting the growth rate of some economic variable W_t such as GDP, the price level or employment. The growth rate for W_t is defined as:

$$\Delta X_t = \ln(W_t) - \ln(W_{t-1}). \tag{4.69}$$

The optimal forecast of ΔX_{t+k} is then $E_t[\Delta X_{t+k}]$ while a 95% confidence interval will take the form:

$$E_t[\Delta X_{t+k}] \pm 1.96\sqrt{Var_t[\Delta X_{t+k}]}. \tag{4.70}$$

The problem then is to calculate $E_t[\Delta X_{t+k}]$ and $Var_t[\Delta X_{t+k}]$.

In the real world instead of ΔX_t the growth rate of W_t is usually defined as g_t where g_t is given by:

$$g_t = \frac{W_t - W_{t-1}}{W_{t-1}}.$$

In practice it generally makes no real practical difference whether g_t or ΔX_t is used. If for example $g_t = 0.04$ or 4% growth, then ΔX_t would be $e^{0.04} - 1 = 0.0408$ or 4.08% growth. In any case it is easy to convert from g_t to ΔX_t using

$$g_t = e^{\Delta X_t} - 1$$

or from ΔX_t to g_t using:

$$\Delta X_t = \ln(1 + g_t).$$

Furthermore if (4.70) is a 95% confidence interval for ΔX_{t+k} then an equivalent 95% confidence interval for g_{t+k} would be:

$$\Pr[b_1 \le g_{t+k} \le b_2] = 0.95$$

where:

$$\begin{aligned} b_1 &= \exp\left(E_t[\Delta X_{t+k}] - 1.96\sqrt{Var_t[\Delta X_{t+k}]}\right) - 1 \\ b_2 &= \exp\left(E_t[\Delta X_{t+k}] + 1.96\sqrt{Var_t[\Delta X_{t+k}]}\right) - 1. \end{aligned}$$

Let us quickly review forecasting Y_t. We now know how to forecast and construct confidence intervals for Y_t defined by:

$$Y_t = \ln(W_t) - (\alpha + \mu t) \tag{4.71}$$

for TS models and

$$Y_t = \ln(W_t) - \ln(W_{t-1}) - \mu \tag{4.72}$$

for DS models. From Theorem 146 we can recursively calculate forecasts of Y_{t+k} as:

$$E_t[Y_{t+k}] = \sum_{j=1}^{p} \phi_j E_t[Y_{t+k-j}] + \sum_{j=1}^{q} \theta_j E_t[a_{t+k-j}] \tag{4.73}$$

where $E_t[Y_{t+k}] = Y_{t+k}$ and $E_t[a_{t+k}] = a_{t+k}$ for $k \leq 0$. Confidence intervals for Y_{t+k} can then be calculated using Theorem 151 using

$$Var_t[Y_{t+k}] = \sigma^2 \sum_{j=0}^{k-1} \psi_j^2 \tag{4.74}$$

where ψ_k is recursively calculated from:

$$\begin{aligned} \psi_k &= \sum_{j=1}^{q} \phi_j \psi_{k-j} + \theta_k \text{ for } k = 0, 1, 2, \ldots q \\ &= \sum_{j=1}^{q} \phi_j \psi_{k-j} \text{ for } k > q \end{aligned} \tag{4.75}$$

with $\psi_0 = 1$ and $\psi_k = 0$ for $k < 0$.

Forecasting ΔX_{t+k} from a DS Model

Let us first consider the case of forecasting for the DS model since this is quite easy. We have:

Theorem 153 *If X_t is DS then the optimal forecast for ΔX_{t+k} is:*

$$E_t[\Delta X_{t+k}] = E_t[Y_{t+k}] + \mu$$

with:

$$Var_t[\Delta X_{t+k}] = \sigma^2 \sum_{j=0}^{k-1} \psi_j^2$$

and a 95% confidence interval for ΔX_{t+k} is:

$$(E_t[Y_{t+k}] + \mu) \pm 1.96\sigma \sqrt{\sum_{j=0}^{k-1} \psi_j^2}. \tag{4.76}$$

Proof. Since for a DS model:

$$\Delta X_{t+k} = \mu + Y_{t+k}$$

it follows that:

$$E_t\left[\Delta X_{t+k}\right] = \mu + E_t\left[Y_{t+k}\right].$$

Furthermore, since μ is a constant:

$$\begin{aligned} Var_t\left[\Delta X_{t+k}\right] &= Var_t\left[\mu + Y_{t+k}\right] \\ &= Var_t\left[Y_{t+k}\right] \\ &= \sigma^2 \sum_{j=0}^{k-1} \psi_j^2. \end{aligned}$$

■

Forecasting ΔX_{t+k} from a TS Model

Now consider the problem of forecasting ΔX_{t+k} from the TS model. Since:

$$X_t = \ln\left(W_t\right) = \alpha + \mu t + Y_t$$

we now have:

$$\Delta X_t \equiv \ln\left(W_t\right) - \ln\left(W_{t-1}\right) = Y_t - Y_{t-1} + \mu. \tag{4.77}$$

To construct a confidence interval for ΔX_{t+k} we then have:

Theorem 154 *If X_t is TS then the optimal forecast for ΔX_{t+k} is:*

$$E_t\left[\Delta X_{t+k}\right] = E_t\left[Y_{t+k}\right] - E_t\left[Y_{t+k-1}\right] + \mu$$

with:

$$\begin{aligned} Var_t\left[\Delta X_{t+k}\right] &= \sigma^2 \sum_{j=0}^{k-1} \left(\psi_j - \psi_{j-1}\right)^2 \\ &= \sigma^2 \left(1 + \sum_{j=1}^{k-1} \left(\psi_j - \psi_{j-1}\right)^2\right) \end{aligned}$$

and a 95% confidence interval for ΔX_{t+k} is:

$$\left(E_t\left[Y_{t+k}\right] - E_t\left[Y_{t+k-1}\right] + \mu\right) \pm 1.96\sigma \left(\sum_{j=0}^{k-1} \left(\psi_j - \psi_{j-1}\right)^2\right)^{\frac{1}{2}}.$$

Proof. Since:

$$\Delta X_{t+k} = Y_{t+k} - Y_{t+k-1} + \mu \tag{4.78}$$

it follows that:

$$E_t\left[\Delta X_{t+k}\right] = E_t\left[Y_{t+k}\right] - E_t\left[Y_{t+k-1}\right] + \mu. \tag{4.79}$$

From (4.75)

$$\begin{aligned} Y_{t+k} &= \sum_{j=0}^{\infty} \psi_j a_{t+k-j} \\ Y_{t+k-1} &= \sum_{j=0}^{\infty} \psi_j a_{t+k-1-j} \end{aligned} \tag{4.80}$$

so that:

$$\begin{aligned} \Delta X_{t+k} &= Y_{t+k} - Y_{t+k-1} + \mu \\ &= \sum_{j=0}^{\infty} \psi_j a_{t+k-j} - \sum_{j=0}^{\infty} \psi_j a_{t+k-1-j} + \mu \\ &= a_{t+k} + \sum_{j=1}^{\infty} \left(\psi_j - \psi_{j-1}\right) a_{t+k-j} + \mu. \end{aligned} \tag{4.81}$$

Using this and (4.78) it follows that:

$$Var_t\left[\Delta X_{t+k}\right] = \sigma^2 \sum_{j=0}^{k-1} \left(\psi_j - \psi_{j-1}\right)^2.$$

■

Remark 155 *Note that for $j = 0$ the term in the sum*

$$\sum_{j=0}^{k-1} \left(\psi_j - \psi_{j-1}\right)^2$$

becomes $\left(\psi_0 - \psi_{-1}\right)^2 = 1$ since $\psi_{-1} = 0$ and $\psi_0 = 1$.

4.4 Box-Jenkins Identification

4.4.1 Identification using $\rho(k)$ and ϕ_{kk}

Box-Jenkins identification is a method, based on the estimated autocorrelation and partial autocorrelation functions, of determining whether a given series is better described by an AR(p) or an MA(q) model.

To simplify matters let us first suppose we know the actual autocorrelation function $\rho(k)$ and the actual partial autocorrelation function ϕ_{kk}.

If Y_t is an AR(p) then $\rho(k)$ will be a damped exponential while ϕ_{kk} will have a cutoff at $k = p$; that is:

$$\begin{aligned} \phi_{kk} &\neq 0, \text{ for } k = 1, 2, \ldots p \\ &= 0, \text{ for } k = p + 1, p + 2, \ldots \infty. \end{aligned} \tag{4.82}$$

Thus if we observe that $\rho(k)$ is a damped exponential and ϕ_{kk} has a finite cutoff at some value of k, then we know that the series is an AR(p) and furthermore we can determine p from the cutoff value.

For example if we were to observe:

$k =$	1	2	3	4	5	6	7	8	9	10
$\rho(k) =$	0.57	0.38	0.34	0.25	0.19	0.14	0.07	0.03	0.01	0.005
$\phi_{kk} =$	0.57	0.08	0	0	0	0	0	0	0	0

then we know right away that Y_t is an AR(p) and not an MA(q) since $\rho(k)$ behaves like a damped exponential; that is it decays but never reaches zero. Furthermore we know that $p = 2$ since ϕ_{kk} has a cutoff at $k = 2$. Thus Y_t follows an AR(2) .

What if Y_t follows an MA(q) process? Then we know that $\rho(k)$ will have a cutoff at $k = q$. What about ϕ_{kk}? It turns out that if Y_t is an MA(q) then ϕ_{kk} will be a damped exponential. This is the reverse of the case where Y_t is an AR(p).

Thus if we were to observe that:

$k =$	1	2	3	4	5	6	7	8	9	10
$\rho(k) =$	0.17	0.38	0.34	0	0	0	0	0	0	0
$\phi_{kk} =$	0.17	0.36	0.24	0.16	0.08	-0.04	0.02	-0.01	0.005	-0.001

then we know that Y_t is not an AR(p) since ϕ_{kk} is a damped exponential and $\rho(k)$ has a cutoff. Furthermore we know that $q = 3$ since the cutoff for ϕ_{kk} occurs at $k = 3$ so that Y_t is an MA(3).

This then is the essence of the Box-Jenkins identification procedure which can be summarized by the following table:

Model	$\rho(k)$	ϕ_{kk}
AR(p)	Damped Exponential	Cutoff at $k = p$
MA(q)	Cutoff at $k = q$	Damped Exponential

.

4.4.2 Estimating $\rho(k)$ and ϕ_{kk}

In practice we do not observe $\rho(k)$ and ϕ_{kk} but must estimate them from the data. Let us begin by estimating the autocovariance function. Given a sample of T observations the autocovariance function $\gamma(k)$ can be estimated by:

Definition 156 ***Estimated Autocovariance Function:***

$$\hat{\gamma}(k) \equiv \frac{1}{T}\sum_{t=1}^{T-|k|}\left(Y_t - \overline{Y}\right)\left(Y_{t+k} - \overline{Y}\right) \quad \text{for } |k| < T. \tag{4.83}$$

Remark 157 *Quite often it will be the case that $\overline{Y} = 0$; for example with either the TS and DS detrending methods models where Y_t is the least squares residual from a regression with a constant term so that $\overline{Y} = 0$. In such cases or where we know a priori that $E[Y_t] = 0$ that we can use:*

$$\hat{\gamma}(k) \equiv \frac{1}{T}\sum_{t=1}^{T-|k|} Y_t Y_{t+k}. \tag{4.84}$$

From $\hat{\gamma}(k)$ we can construct an estimate of the autocorrelation function as:

Definition 158 ***Estimated of the Autocorrelation Function:***

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}. \tag{4.85}$$

With Box-Jenkins identification we need to standard errors to determine if $\hat{\rho}(k)$ is significantly different than zero. We have:[1]

Proposition 159 *The asymptotic distribution of $\hat{\rho}(k)$ is of the form:*

$$\sqrt{T}\left(\hat{\rho}(k) - \rho(k)\right) \overset{a}{\sim} N[0, V(k)]$$

where $V(k)$ is given by:

$$V(k) = \sum_{j=-\infty}^{\infty}\left(\rho(j)^2 + \rho(j+k)\rho(j-k) + 2\rho(k)^2\rho(j)^2 - 4\rho(k)\rho(j)\rho(j-k)\right). \tag{4.86}$$

Remark 160 *$\hat{\rho}(k)$ and $\hat{\rho}(l)$ will in general be asymptotically correlated with each other with the asymptotic correlation an even more complicated function than $V(k)$.*

If $\rho(k) = 0$ for $k = 1, 2, \ldots \infty$ (i.e. Y_t is white noise) then things simplify considerably. In this case $V(k) = 1$ so that:

$$\sqrt{T}\hat{\rho}(k) \overset{a}{\sim} N[0, 1] \tag{4.87}$$

and the $\hat{\rho}(k)$'s are asymptotically uncorrelated. Although not strictly speaking correct, we often make this assumption when doing Box-Jenkins identification

[1] See Priestly, p332

analysis so that, at least approximately, a standard error for $\hat{\rho}(k)$ is calculated as:

$$SE\left[\hat{\rho}(k)\right] = \frac{1}{\sqrt{T}}. \tag{4.88}$$

Thus using a two-sigma rule,[2] $\hat{\rho}(k)$ is taken as being significantly different than zero if:

$$\left|\hat{\rho}(k)\right| > \frac{2}{\sqrt{T}}. \tag{4.89}$$

For example given $T = 150$ observations we would have:

$$2 \times SE\left[\hat{\rho}(k)\right] = \frac{2}{\sqrt{150}} = 0.16 \tag{4.90}$$

so that $\hat{\rho}(k)\,'s$ greater than about 0.16 would be taken as being significant.

We can estimate the partial autocorrelation function by ϕ_{kk} by replacing $\rho(k)$ with $\hat{\rho}(k)$ in the Yule-Walker equations. These have a nicer asymptotic distribution. In particular

Proposition 161 *If $\phi_{kk} = 0$ for $k > p$ (i.e., Y_t is an AR(p)) then:*

$$\sqrt{T}\hat{\phi}_{kk} \overset{a}{\sim} N\left[0,1\right] \quad \text{for } k > p \tag{4.91}$$

and the $\hat{\phi}_{kk}\,'s$ are asymptotically uncorrelated with each other.

Thus

$$SE\left[\hat{\phi}_{kk}\right] = \frac{1}{\sqrt{T}} \tag{4.92}$$

so that, again using a two-sigma rule, $\hat{\phi}_{kk}$ would be taken as significantly different than zero if:

$$\left|\hat{\phi}_{kk}\right| > \frac{2}{\sqrt{T}}. \tag{4.93}$$

Again given $T = 150$ observations $\frac{2}{\sqrt{T}} = 0.16$ so that $\hat{\phi}_{kk}\,'s$ greater than about 0.16 in absolute value would be taken as significantly different that zero.

It turns out that when $p > 0$ and $q > 0$ that both the autocorrelation function $\rho(k)$ and the partial autocorrelation function ϕ_{kk} behave like damped exponentials so that:

Model	$\rho(k)$	ϕ_{kk}
AR(p)	Damped Exponential	Cutoff at $k = p$
MA(q)	Cutoff at $k = q$	Damped Exponential
ARMA(p,q)	Damped Exponential	Damped Exponential

.

[2] You could use 1.96 from the standard normal tables instead of 2 but this level of precision is really not necessary given the looseness in the way distributional theory is used for Box-Jenkins identification.

This means that it is only when we observe that both $\rho(k)$ and ϕ_{kk} appear as damped exponentials that we should choose a mixed process. Unfortunately if we observe such a pattern there is nothing in either $\rho(k)$ or ϕ_{kk} to tell us what p and q are. This problem is made even more difficult when we use the estimated functions $\hat{\rho}(k)$ and $\hat{\phi}_{kk}$. Although Box and Jenkins discuss methods of identifying mixed models, generally in practice and with commonly found sample sizes, these methods are of little or no value.

In practice this is not such a serious problem since generally either a pure AR(p) or MA(q) will be found which fit the data quite well. Usually then there is no need to consider mixed processes. The only situation when you might is when both a pure AR(p) or MA(q) would involve a large number of parameters.

For example suppose $\hat{\rho}(k)$ has significant values out to $k = 10$ while $\hat{\phi}_{kk}$ has significant values out to $k = 12$. In this case you would be choosing between either an MA(10) or an AR(12) , neither of which is very parsimonious. In this case you might try say an ARMA(2,1) and see if you could get by with estimating 3 parameters instead of 10 or 12.

4.5 Maximum Likelihood Estimation

4.5.1 Calculating the Log-Likelihood

Once we have identified an ARMA(p,q) model we will need to estimate it. We have:

Theorem 162 *For an ARMA(p,q) the log-likelihood is given by:*

$$l\left(\phi,\theta,\sigma^2\right) = -\frac{T}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{T} a_t\left[\phi,\theta\right]^2 \tag{4.94}$$

where $a_t[\phi,\theta]$ is calculated recursively by:

$$a_t\left[\phi,\theta\right] = Y_t - \sum_{j=1}^{p}\phi_j Y_{t-j} - \sum_{j=1}^{q}\theta_j a_{t-j}\left[\phi,\theta\right]. \tag{4.95}$$

Proof. From the ARMA(p,q) model:

$$Y_t = \sum_{j=1}^{p}\phi_j Y_{t-j} + a_t + \sum_{j=1}^{q}\theta_j a_{t-j} \tag{4.96}$$

define the conditional expectation of Y_t given the information set at time $t-1$, the AR parameters ϕ and the MA parameters θ as:

$$\begin{aligned} E_{t-1}\left[Y_t|\phi,\theta\right] &= \sum_{j=1}^{p}\phi_j Y_{t-j} + \sum_{j=1}^{q}\theta_j a_{t-j}\left[\phi,\theta\right] && (4.97)\\ a_t\left[\phi,\theta\right] &= Y_t - E_{t-1}\left[Y_t|\phi,\theta\right]. && (4.98)\end{aligned}$$

Then

$$Y_t | I_{t-1} \sim N\left[E_{t-1}\left[Y_t | \phi, \theta\right], \sigma^2\right]$$

so that:

$$\begin{aligned} p(Y_t | I_{t-1}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(Y_t - E_{t-1}\left[Y_t|\phi,\theta\right])^2}{\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{a_t\left[\phi,\theta\right]^2}{\sigma^2}\right). \end{aligned}$$

Following the derivation of the likelihood for the AR(p) process in Section 3.8.2 the result follows. ■

In order to calculate $a_t\left[\phi,\theta\right]$ from the recursive formula in (4.95) one needs starting values. One way of doing this is to use the unconditional mean 0 so that:

$$a_t\left[\phi,\theta\right] = E\left[a_t\right] = 0, \text{ for } t \leq 0.$$

As long as the chosen model is invertible and the number of observations is reasonably large the choice of these starting values will have little effect on the ML estimates. Box and Jenkins suggest backcasting as a method of obtaining better starting values.

4.5.2 Estimating ϕ and θ

To estimate ϕ and θ we need to minimize the sum of squares:

$$S(\phi,\theta) = \sum_{t=1}^{T} a_t\left[\phi,\theta\right]^2. \tag{4.99}$$

The maximum likelihood estimators: $\hat{\phi}, \hat{\theta}$ are then the solutions to the first-order conditions:

$$\frac{\partial S\left(\hat{\phi},\hat{\theta}\right)}{\partial \phi} = 0, \frac{\partial S\left(\hat{\phi},\hat{\theta}\right)}{\partial \theta} = 0.$$

It $q > 0$ this will require a nonlinear optimization procedure such as Newton's method where consistent starting values $\hat{\phi}_0$, $\hat{\theta}_0$ are chosen and one iterates as:

$$\begin{bmatrix} \hat{\phi}_n \\ \hat{\theta}_n \end{bmatrix} = \begin{bmatrix} \hat{\phi}_{n-1} \\ \hat{\theta}_{n-1} \end{bmatrix} - \begin{bmatrix} \frac{\partial^2 S\left(\hat{\phi}_{n-1},\hat{\theta}_{n-1}\right)}{\partial\phi\partial\phi^T} & \frac{\partial^2 S\left(\hat{\phi}_{n-1},\hat{\theta}_{n-1}\right)}{\partial\theta\partial\phi^T} \\ \frac{\partial^2 S\left(\hat{\phi}_{n-1},\hat{\theta}_{n-1}\right)}{\partial\phi\partial\theta^T} & \frac{\partial^2 S\left(\hat{\phi}_{n-1},\hat{\theta}_{n-1}\right)}{\partial\theta\partial\theta^T} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial S\left(\hat{\phi}_{n-1},\hat{\theta}_{n-1}\right)}{\partial\phi} \\ \frac{\partial S\left(\hat{\phi}_{n-1},\hat{\theta}_{n-1}\right)}{\partial\theta} \end{bmatrix}$$

until convergence takes place. Such nonlinear optimization procedures are now standard in econometric software such as TSP, $GAUSS$, and $RATS$.

4.5.3 The MA(1) Model

For an MA(1) we have:

$$\begin{aligned} a_t[\theta] &= Y_t - \theta a_{t-1}[\theta] \\ &= Y_t - \theta\left(Y_{t-1} - \theta a_{t-2}[\theta]\right) \\ &\vdots \\ &= Y_t - \theta Y_{t-1} + \theta^2 Y_{t-2} - \theta^3 Y_{t-3} + \cdots + (-\theta)^{t-1} Y_t + (-\theta)^t a_0[\theta]. \end{aligned} \tag{4.100}$$

As long as the process is invertible or $|\theta| < 1$, the choice of $a_0[\theta]$ will have little impact on most of the $a_t[\theta]' s$ since $(-\theta)^t$ will generally be a small number. For example if $\theta = 0.5$ and $t = 40$ we would have as a coefficient on $a_0[\theta]$:

$$(-0.5)^{40} = 0.000000000000909. \tag{4.101}$$

If we set the starting values $a_0[\theta] = 0$, then we obtain:

$$a_t[\theta] = Y_t - \sum_{j=1}^{t-1} (-1)^{j-1} \theta^j Y_{t-j}. \tag{4.102}$$

Note that unlike the AR model this is a *nonlinear* function of θ since we have the powers: θ^j. This means that when we minimize the sum of squares:

$$S(\theta) = \sum_{t=1}^{T} \left(Y_t - \sum_{j=1}^{t-1} (-1)^{j-1} \theta^j Y_{t-j} \right)^2. \tag{4.103}$$

that minimizing $S(\theta)$ requires an iterative quadratic hill-climbing procedure such as Newton's method.

Once $\hat{\theta}$ has been found, however, things proceed as before with the AR(p) model. Define the least squares residual $\hat{a}_t$ by:

$$\hat{a}_t = a_t\left[\hat{\theta}\right] = Y_t - \sum_{j=1}^{q} \hat{\theta}_j \hat{a}_{t-j}. \tag{4.104}$$

Once $\hat{\theta}$ and $\hat{\phi}$ are found by such a non-linear procedure things become easy again and we can proceed as before. We have:

Theorem 163 *The ML estimator of σ^2 is given by:*

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \hat{a}_t^2 \tag{4.105}$$

where:

$$\hat{a}_t = a_t\left[\hat{\phi}, \hat{\theta}\right] \tag{4.106}$$

and the maximized log-likelihood is:

$$l^* \equiv l\left(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2\right) = -\frac{T}{2} \ln\left(\hat{\sigma}^2\right) - \frac{T}{2}. \tag{4.107}$$

Proof. Solving for the maximum likelihood estimator of $\hat{\sigma}^2$ from the first-order condition

$$\frac{\partial l\left(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2\right)}{\partial \sigma^2} = -\frac{T}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2\left(\hat{\sigma}^2\right)^2}\sum_{t=1}^{T}\hat{a}_t^2 = 0$$

we obtain:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{a}_t^2. \tag{4.108}$$

The proof for l^* is the same as for the AR(p) model in Section 3.8.2. ■

4.5.4 Likelihood Ratio Tests

Since the formula in (4.107) is identical to that for the AR(p) model, likelihood ratio tests are the same as in Section 3.8.2. We proceed by estimating a restricted and an unrestricted version of an ARMA process. The restricted and unrestricted maximized log-likelihoods are then:

$$\begin{aligned} l_R^* &= -\frac{T}{2}\ln\left(\hat{\sigma}_R^2\right) - \frac{T}{2} \text{ and} \qquad (4.109)\\ l_U^* &= -\frac{T}{2}\ln\left(\hat{\sigma}_U^2\right) - \frac{T}{2} \end{aligned}$$

where $\hat{\sigma}_R^2$ and $\hat{\sigma}_U^2$ are the restricted estimator and unrestricted estimators of σ^2. The likelihood ratio statistic:

$$\Lambda = -2\left(l_R^* - l_U^*\right) = T\ln\left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2}\right) \tag{4.110}$$

then has under the null that the restrictions are true an asymptotic distribution:

$$\Lambda \overset{a}{\sim} \chi_r^2 \tag{4.111}$$

where r is the number of restrictions or the difference in the number of parameters in the restricted and unrestricted models.

For example suppose we believe that Y_t in an MA(2) but wish to apply a diagnostic test to see whether this belief is consistent with the data. To do this we overfit by estimating an MA(5):

$$Y_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \theta_3 a_{t-3} + \theta_4 a_{t-4} + \theta_5 a_{t-5}.$$

If our belief of an MA(2) is correct then $\theta_3 = \theta_4 = \theta_5 = 0$. We therefore test:

$$\begin{aligned} H_0 &: Y_t \sim \text{MA}(2) \ \ (\text{or } \theta_3 = \theta_4 = \theta_5 = 0) \ \ \text{versus} \\ H_1 &: Y_t \sim \text{MA}(5) \ (\text{or } \theta_3 \neq 0 \text{ or } \theta_4 \neq 0 \text{ or } \theta_5 \neq 0). \end{aligned}$$

Suppose we have $T = 100$ observations and for the restricted MA(2) model we find that:

$$\hat{\sigma}_R^2 = \frac{(0.15)^2}{100}$$

where $(0.15)^2$ is the restricted sum of squares. We then estimate the unrestricted MA(5) and find

$$\hat{\sigma}_U^2 = \frac{(0.11)^2}{100}$$

where $(0.11)^2$ is the unrestricted sum of squares.

We then have:

$$\Lambda = 100 \ln \left(\frac{\frac{(0.15)^2}{100}}{\frac{(0.11)^2}{100}} \right) = 100 \ln \left(\frac{(0.15)^2}{(0.11)^2} \right) = 62.03 > \chi_3^2(0.05) = 7.815 \quad (4.112)$$

and so we *reject* H_0 that Y_t is an $MA(2)$.

4.5.5 Estimating p and q

We can also calculate either the Akaike or Schwarz criteria for an ARMA(k,l) as:

$$AIC(k,l) = \ln\left(\hat{\sigma}_{k,l}^2\right) + \frac{2(k+l)}{T} \quad (4.113)$$

$$SC(k,l) = \ln\left(\hat{\sigma}_{k,l}^2\right) + \frac{\ln(T)(k+l)}{T}$$

for $k = 0, 1, 2, \dots p_{\max}$ and $l = 0, 1, 2, \dots q_{\max}$ where k is the order of the AR and l the order of the MA with $\hat{\sigma}_{k,l}^2$ the estimated value of σ^2 for an ARMA(k,l) process.

Thus in principle we could use either the Akaike or Schwarz criterion to estimate p and q. This would seem to get around the problem of determining p and q for mixed processes associated with the Box-Jenkins identification procedure. A number of difficulties should be noted however. First when $l > 0$ we require an iterative procedure to estimate the ARMA(k,l) process and for this we will need to worry about convergence and starting values which may make it very time consuming to implement in practice. Secondly, we might have difficulties with near cancellation of common roots for many values of k and l. Thus for example if an ARMA(1,1) represents the data well, we are likely to find numerical problems if we try and estimate an ARMA(2,4).

4.6 ARIMA(p,d,q) models

Box and Jenkins proposed a general class of models called ARIMA models (autoregressive integrated moving average processes) defined as:

Definition 164 *We say that $X_t \sim$ ARIMA(p,d,q) if :*

$$\phi(B)(1-B)^d X_t = \alpha + \theta(B) a_t.$$

where p is the order of the AR polynomial $\phi(B)$, q is the order of the MA polynomial $\theta(B)$, and d is the number of times that X_t must be differenced to achieve stationarity.

In economics the only values of d which generally make sense are

$$d = 0,\ 1,\ 2.$$

A value of $d = 0$ corresponds to the case where X_t is already stationary and so does not require differencing. Examples of time series where $d = 0$ is reasonable would be the growth rate of real GDP, the rate of unemployment or the real rate of interest. In this case

$$\mu = E[X_t] = \frac{\alpha}{\phi(1)}$$

is the mean of the series.

The case of $d = 1$ corresponds to the case where X_t is not stationary but $(1-B)X_t$ is stationary. This is essentially the DS model where:

$$Y_t = (1-B)X_t - \mu \tag{4.114}$$

with growth rate given by:

$$\mu = E[(1-B)X_t] = \frac{\alpha}{\phi(1)}$$

and where Y_t follows a stationary ARMA(p,q) process with mean 0.

It is hard to think of many examples where $d = 2$ might make economic sense since this implies that the growth rate of the series is nonstationary. One possible example is if $X_t = \ln(P_t)$ where P_t is the CPI or the GDP price deflator. In this case $(1-B)X_t$ would be the rate of inflation. Given the recent inflationary history of many modern economies, it is not unreasonable to assume that inflation is nonstationary and hence would require differencing so that only $(1-B)^2 X_t$, or the change in inflation, is stationary. Such a model then would be of the form:

$$\phi(B)(1-B)^2 X_t = \alpha + \theta(B) a_t \tag{4.115}$$

in which case

$$\mu = E\left[(1-B)^2 X_t\right] = \frac{\alpha}{\phi(1)}$$

would be the average increase in the rate of inflation. Unless the economy is experiencing hyperinflation it would probably make sense to set $\alpha = 0$.

4.6.1 Box Jenkins Identification

Generally speaking economic considerations pretty well determine the value of d although there are some exceptions. For example many economic theories predict the real rate of interest should be stationary even though actual real rates of interest often appear to be nonstationary.

Box and Jenkins suggest an examination of the estimated autocorrelation function $\hat{\rho}(k)$ and partial autocorrelation function $\hat{\phi}_{kk}$ in order to determine d. If a series is stationary both $\hat{\rho}(k)$ and $\hat{\phi}_{kk}$ should have the short memory property and so decay rapidly to zero. We have:

Proposition 165 *For a nonstationary series that needs differencing the typical pattern one sees is that $\hat{\rho}(k)$ decays* linearly *(typically with $\hat{\rho}(1) \approx 1$) while $\hat{\phi}_{kk}$ has one big spike at $k = 1$ (typically with $\hat{\phi}_{11} \approx 1$) with $\hat{\phi}_{kk} \approx 0$ for $k > 1$.*

The linear decay in particular is inconsistent with the short-memory property which requires *exponential* rather than linear decay.

Thus the approach suggested by Box and Jenkins is to continue differencing the series until $\hat{\rho}(k)$ and $\hat{\phi}_{kk}$ are consistent with exponential decay.

Once stationarity has been achieved it is important *not* to overdifference. A value of $\hat{\rho}(1) \approx -0.5$ is often indicative of over-differencing. For example if Y_t is stationary white noise so that: $Y_t = a_t$ but you incorrectly difference Y_t then the differenced series $\widetilde{Y}_t \equiv (1 - B) Y_t$ will be a non-invertible MA(1) since:

$$\widetilde{Y}_t = a_t - a_{t-1} \tag{4.116}$$

In this case $\rho(1) = -\frac{1}{2}$ since for an MA(1) with $\theta = -1$ we have:

$$\rho(1) = \frac{\theta}{1 + \theta^2} = -\frac{1}{2}. \tag{4.117}$$

4.7 Diagnostic Tests

4.7.1 Introduction

By the time we have arrived at our chosen ARMA(p,q) model many decisions have had to be made. Should I detrend using TS or DS? Do I need to correct for seasonality and if so, is it better to use seasonal dummies or seasonal differencing? There was one significant autocorrelation at lag 7 which I ignored (or did not ignore) was that correct?

Each time we make a decision there arises the possibility of error. The purpose of diagnostic testing then is to make us aware of errors that we have made so that we have a chance to correct them.

4.7.2 Catastrophic Errors

The worst kind of errors are catastrophic errors. Catastrophic errors are errors which mean that your results are garbage; catastrophic errors make you look foolish at best and may mean losing your job at worst.

Unfortunately catastrophic errors often occur innocently, in a manner entirely disproportionate to their effects. For example, the data in the file you are using contain two columns, in the first is the date of each observation while in the second is say GDP; that is something like this:

1956	478.2
1957	483.9
1958	492.1
1959	498.7
etc.	etc..

You don't realize that the date is in the first column and ask the computer to read the data from this file. Computers being what they are do this without complaining so the GDP series is loaded into the computer as:

1956 478.2 1957 483.9 1958 etc..

You then go ahead and do unit root tests and identify an ARMA(p,q) model for the series.

Your results are, unfortunately, junk. Your data has GDP in 1956 being equal to 1956 with a massive recession in 1957 when GNP falls to 478.2; a 76% fall in output; something only a nuclear war would be likely to accomplish. You have made a catastrophic error.

It is not a problem to make this sort of error, in fact you will make this sort of error regularly if you do applied work. The important point is not to let it slip by you without noticing it; not, for example, to hand it in as an assignment or give it to your boss in your annual report.

If you do make this sort of error without catching it first, it is often because you have never really looked at your data or at what you have given the computer. You are acting like a machine. Catastrophic errors are characteristic of an age where computers allow you to perform all sorts of sophisticated statistical analysis without ever actually looking at the data. As the old expression goes, garbage in, garbage out.

Avoiding catastrophic errors is often quite easy. The two essential commands are PRINT and PLOT. If you get in the habit of printing out and plotting your data and intermediate results, it is much less likely that you miss a catastrophic error. The PRINT and PLOT commands are therefore your most important diagnostic tests; to be performed before any other. Use them regularly!

4.7.3 Regular Errors

In applied work you can safely bet that the model you choose is not going to be the true model; whatever "true model" means. More importantly, the model you have chosen may not be adequate over certain dimensions.

There are now a huge number of diagnostic tests available in the literature and in many computer programs. We will concentrate on the following:

1. Overfitting ,

2. Tests for Normality,
3. Tests for White Noise and
4. A Lagrange Multiplier Test for ARCH.

4.7.4 Overfitting

Suppose your chosen model is ARMA(p,q):

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + a_t + \sum_{j=1}^{q} \theta_j a_{t-j} \tag{4.118}$$

but you think that maybe p or q should have been larger.

Assuming the model in (4.118) is true it follows that it is also an ARMA(p+r,q) model where:

$$\begin{aligned} Y_t &= \sum_{j=1}^{p} \phi_j Y_{t-j} + a_t + \sum_{j=1}^{q} \theta_j a_{t-j} \\ &\quad + \phi_{p+1} Y_{t-(p+1)} + \phi_{p+2} Y_{t-(p+2)} + \cdots + \phi_{p+r} Y_{t-(p+r)} \end{aligned} \tag{4.119}$$

and where:

$$\phi_{p+1} = \phi_{p+2} = \cdots = \phi_{p+r} = 0. \tag{4.120}$$

Alternatively if (4.118) is true then Y_t also follows an ARMA(p,q+r) model as:

$$\begin{aligned} Y_t &= \sum_{j=1}^{p} \phi_j Y_{t-j} + a_t + \sum_{j=1}^{q} \theta_j a_{t-j} \\ &\quad + \theta_{q+1} a_{t-(q+1)} + \theta_{q+2} a_{t-(q+2)} + \cdots + \theta_{q+r} a_{t-(q+r)} \end{aligned} \tag{4.121}$$

where:

$$\theta_{q+1} = \theta_{q+2} = \cdots = \theta_{q+r} = 0. \tag{4.122}$$

If however the chosen ARMA(p,q) in (4.118) is false or inadequate, then we would expect the restrictions in either (4.120) or (4.122) to be rejected by the data.[3]

This then is the basis for overfitting as a diagnostic test: we estimate a larger model that includes our model as a special case and see if the data is consistent with the implied restrictions. Thus if we make the ARMA(p,q) model

[3] You should not increase both p and q at the same time since under the null hypothesis of an $ARMA\,(p,q)$ the parameters of an $ARMA\,(p+r,q+r)$ are not identified due to the cancellation of common factors.

or (4.120) as the null hypothesis and an ARMA(p+r,q) as the alternative then the maximized log-likelihood of the unrestricted model is:

$$l^*_{p+r,q} = -\frac{T}{2}\ln\left(\hat{\sigma}^2_{p+r,q}\right) - \frac{T}{2} \tag{4.123}$$

where $\hat{\sigma}^2_{p+r,q}$ is the estimator of σ^2 from the unrestricted ARMA(p+r,q) model. Similarly the maximized log-likelihood of the restricted model is:

$$l^*_{p,q} = -\frac{T}{2}\ln\left(\hat{\sigma}^2_{p,q}\right) - \frac{T}{2}$$

where $\hat{\sigma}^2_{p,q}$ is the estimator of σ^2 from the restricted ARMA(p,q) model. We then have under the null hypothesis that the ARMA(p,q) model is adequate that for the likelihood ratio statistic:

$$\begin{aligned} \Lambda &= -2\left(l^*_R - l^*_U\right) \\ &= T\ln\left(\frac{\hat{\sigma}^2_{p,q}}{\hat{\sigma}^2_{p+r,q}}\right) \overset{a}{\sim} \chi^2_r. \end{aligned} \tag{4.124}$$

Similarly if the alternative is an ARMA(p,q+r) then we would use

$$\Lambda = T\ln\left(\frac{\hat{\sigma}^2_{p,q}}{\hat{\sigma}^2_{p,q+r}}\right) \overset{a}{\sim} \chi^2_r.$$

4.7.5 Tests for Normality

If we estimate our ARMA(p,q) model as:

$$\hat{\phi}(B)\,Y_t = \hat{\theta}(B)\,a_t \tag{4.125}$$

and obtain the residuals:

$$\hat{a}_t = \frac{\hat{\phi}(B)}{\hat{\theta}(B)}Y_t \tag{4.126}$$

then if the model is correct $\hat{a}_t$ should be approximately $i.i.n\left(0,\sigma^2\right)$ or alternatively

$$\hat{z}_t = \frac{\hat{a}_t}{\hat{\sigma}} \tag{4.127}$$

should be $i.i.n\,(0,1)$; that is a sequence of independent standard normals.

One very simple but often very effective test for normality is simply to plot $\hat{z}_t$. We know from the properties of the standard normal distribution that if

$$z_t \sim N\,[0,1]$$

then:

$$\begin{aligned} \Pr\left[|z_t| \geq 2\right] &\approx 0.05 \text{ (i.e., one in 20)} \\ \Pr\left[|z_t| \geq 3\right] &\approx 0.0027 \text{ (i.e., one in 370)} \\ \Pr\left[|z_t| \geq 4\right] &\approx 0.00006 \text{ (i.e., one in 15787)} \\ \Pr\left[|z_t| \geq 5\right] &\approx 0.0000006 \text{ (i.e., one in } 1,744,277). \end{aligned} \quad (4.128)$$

It is clear that the tails of the normal distribution rapidly thin out; for example it is almost impossible (in a normal sized sample) that the normal distribution would every produce a z_t greater than 4 in absolute value.

In a typical sample of say $T \approx 150$, we would expect to see about 7 values of $\hat{z}_t$ greater than 2 in absolute value, maybe 1 or 2 greater than 3 in absolute value and none greater than 4 in absolute value. If we do observe even one observation greater than 4, or say ten observations greater than 3, then this by itself is pretty conclusive evidence against the normal distribution.

Sometimes very large values of $\hat{z}_t$ (or outliers) correspond to special historical circumstances: wars, strikes or natural disasters. Therefore it is important to consider when the outliers take place. You might then want to remove them from the sample or include dummy variables when doing the DS or TS detrending.

Real big outliers are often the result of some kind of computing error or recording error or catastrophic error. If you have missed a decimal place in say the third quarter of 1968, typing 96784 instead of 967.84, then this will likely show up a large value of $\hat{z}_t$ in 1968.

We can also test for normality more formally. The two telltale characteristics of the standard normal distribution are symmetry and thin tails which can be measured by skewness and kurtosis. In particular if z_t is a standard normal then:

$$\begin{aligned} \kappa_3 &\equiv E\left[z_t^3\right] = 0 \\ \kappa_4 &\equiv E\left[z_t^4\right] = 3 \end{aligned} \quad (4.129)$$

where κ_3 is the skewness and κ_4 the kurtosis. Two important ways that the actual distribution could differ from the normal is either being skewed (*i.e.* not symmetric) so that $\kappa_3 \neq 0$ or by having thicker tails than the normal distribution so that $\kappa_4 > 3$.

To test normality consider calculating the sample skewness and kurtosis estimates given by:

$$\begin{aligned} \hat{\kappa}_3 &= \frac{1}{T}\sum_{t=1}^{T} \hat{z}_t = \frac{1}{T}\frac{\sum_{t=1}^{T} \hat{a}_t^2}{\hat{\sigma}^4} \\ \hat{\kappa}_4 &= \frac{1}{T}\sum_{t=1}^{T} \hat{z}_t = \frac{1}{T}\frac{\sum_{t=1}^{T} \hat{a}_t^4}{\hat{\sigma}^4}. \end{aligned} \quad (4.130)$$

It can then be shown that under

$$H_0 : \kappa_3 = 0$$

that:

$$\hat{t}_3 = \sqrt{\frac{T}{6}}\hat{\kappa}_3 \overset{a}{\sim} N\left[0,1\right]. \tag{4.131}$$

Similarly given:

$$H_o : \kappa_4 = 3$$

it can be shown that:

$$\hat{t}_4 = \sqrt{\frac{T}{24}}\left(\hat{\kappa}_4 - 3\right) \overset{a}{\sim} N\left[0,1\right]. \tag{4.132}$$

Furthermore $\hat{t}_3$ and $\hat{t}_4$ are asymptotically independent.

We would therefore reject normality because of skewness (at the 5% level) if:

$$\left|\hat{t}_3\right| > 1.96 \tag{4.133}$$

or because of kurtosis if:

$$\left|\hat{t}_4\right| > 1.96. \tag{4.134}$$

If $\hat{t}_3 > 1.96$ then the distribution is skewed to the right ($\hat{z}_t$ probably has more or larger positive outliers) while if $\hat{t}_3 < -1.96$ then the distribution is skewed to the left ($\hat{z}_t$ probably has more or larger negative outliers).

Similarly if $\hat{t}_4 > 1.96$ then the tails of the distribution are significantly thicker than the normal (there will be more large values of $\hat{z}_t$ than the normal distribution can account for) which is while if $\hat{t}_4 < -1.96$ then the tails of the distribution are significantly thinner than the normal distribution (the $\hat{z}_t$'s have fewer outliers than one would expect with the normal distribution). In economics rejection because $\hat{t}_4 > 1.96$ is much more common.

A joint test of zero skewness and excess kurtosis is also available. Under the null of normality or:

$$H_0 : \kappa_3 = 0,\ \kappa_4 = 3$$

we have:

$$JB = \hat{t}_3^2 + \hat{t}_4^2 = T\left(\frac{\hat{\kappa}_3^2}{6} + \frac{\left(\hat{\kappa}_4 - 3\right)^2}{24}\right) \overset{a}{\sim} \chi_2^2. \tag{4.135}$$

This is often referred as the Jarque-Bera test for normality. Thus at the 5% level if $JB > 6$ (the exact critical value is 5.991465) we would reject the null of normality.

4.7.6 Portmanteau Tests of Zero Correlation

Let us assume for the moment that we know the ARMA(p,q) parameters of our model. This would allow us to calculate a_t directly as:

$$a_t = Y_t - \sum_{j=1}^{p} \phi_j Y_{t-j} - \sum_{j=1}^{q} \theta_j a_{t-j}. \tag{4.136}$$

If we have chosen $\phi(B)$ and $\theta(B)$ properly then the series a_t should be *i.i.d.* and hence uncorrelated. From this it follows that the autocorrelation function for a_t defined by:

$$\rho_a(k) = \frac{E[a_t a_{t-k}]}{\sigma^2} \tag{4.137}$$

satisfies:

$$\rho_a(k) = 0 \text{ for } k \neq 0 \tag{4.138}$$

We can test this implication of our model by estimating $\rho_a(k)$ by:

$$\hat{\rho}_a(k) = \frac{\sum_{t=1}^{T-|k|} a_t a_{t-k}}{\sum_{t=1}^{T} a_t^2}. \tag{4.139}$$

Under

$$H_0 : \rho_a(k) = 0 \text{ for } k = 1, 2, \ldots M$$

we have:

$$\sqrt{T}\hat{\rho}_a(k) \overset{a}{\sim} N[0,1]$$

and that $\hat{\rho}_a(k)$ and $\hat{\rho}_a(l)$ are independent for $k \neq l$.

This suggests calculating $\hat{\rho}_a(k)$ for $k = 1, 2, \ldots M$ and comparing them with the standard error $SE[\hat{\rho}_a(k)] = \frac{1}{\sqrt{T}}$ so that if:

$$|\hat{\rho}_a(k)| > \frac{1.96}{\sqrt{T}} \tag{4.140}$$

we would suspect that the *i.i.d.* assumption is not correct and that we haven't chosen $\phi(B)$ or $\theta(B)$ properly.

We can also construct a joint test since under H_0 :

$$Q = T \sum_{k=1}^{M} \hat{\rho}_a(k)^2 \overset{a}{\sim} \chi_M^2 \tag{4.141}$$

which is the Box-Pierce portmanteau test applied to a_t.

Unfortunately these procedures are not feasible since we do not know in practice what are the true parameters of $\phi(B)$ and $\theta(B)$. We can however use the estimated parameters and residuals as:

$$\hat{a}_t = Y_t - \sum_{j=1}^{p} \hat{\phi}_j Y_{t-j} - \sum_{j=1}^{p} \hat{\theta}_j \hat{a}_{t-j} \tag{4.142}$$

and calculate

$$\hat{\rho}_{\hat{a}}(k) = \frac{\sum_{t=1}^{T-|k|} \hat{a}_t \hat{a}_{t-k}}{\sum_{t=1}^{T} \hat{a}_t^2}. \tag{4.143}$$

A complication arises because the noise in the parameter estimates affects the distribution of $\hat{\rho}_{\hat{a}}(k)$, even asymptotically. However, if we let:

$$M \approx \sqrt{T} \tag{4.144}$$

so that M grows as the square root of the number of observations T then for

$$Q = T \sum_{k=1}^{M} \hat{\rho}_{\hat{a}}(k)^2 \tag{4.145}$$

we have:

$$Q \overset{a}{\sim} \chi_M^2 \tag{4.146}$$

This then is the feasible form of the Box-Pierce portmanteau test. Often you will see a small-sample degrees of freedom correction for the asymptotic distribution:

$$Q \overset{a}{\sim} \chi_{M-p-q}^2 \tag{4.147}$$

where p and q are the order of the AR and MA polynomials.

Monte Carlo and other theoretical studies have shown that the Box-Pierce test often does not perform well in a small samples. For this reason a modification is often made known as the Ljung-Box portmanteau test:

$$Q^* = T(T+2) \sum_{k=1}^{M} (T-k)^{-1} \hat{\rho}_{\hat{a}}(k)^2 \tag{4.148}$$

which has the same asymptotic distribution as Q; that is:

$$Q^* \overset{a}{\sim} \chi_M^2 \tag{4.149}$$

but has better small sample characteristics.

Remark 166 *It is important to realize that $\hat{\rho}_a(k)$ is* not *the same as the $\hat{\rho}(k)$ we encountered with Box-Jenkins identification, which is calculated for Y_t and not $\hat{a}_t$; that is:*

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{T-|k|} Y_t Y_{t+k}}{\sum_{t=1}^{T} Y_t^2} \tag{4.150}$$

which is not the same as (4.143). *If the model is correct then we would expect $\hat{\rho}_a(k) \approx 0$ but not that $\hat{\rho}(k) \approx 0$.*

4.7.7 Testing for ARCH

Even if the data appear to be consistent with the a_t *'s* being uncorrelated, it does not follow that the a_t *'s* are independent. This is because zero correlation does not imply independence. For example it even if a_t is uncorrelated with a_{t-1}, it may still be that a_t^2 is correlated with a_{t-1}^2. This is in fact quite common in economics, especially with financial data which experience occasional bursts of volatility.

Autoregressive, conditional, heteroskedastic models of order q or $ARCH(q)$ models display this kind of behavior. We have:

Definition 167 ARCH (q) : *a_t follows an $ARCH(q)$ process if*

$$a_t = z_t \left(\sigma^2 + \alpha_1 a_{t-1}^2 + \alpha_2 a_{t-2}^2 + \cdots + \alpha_q a_{t-q}^2\right)^{1/2}$$

where $z_t \sim i.i.n\,(0,1)$ is a sequence of independent standard normals.

Under the null hypothesis that:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_q = 0 \tag{4.151}$$

the $ARCH$ model reduces to

$$a_t = \sigma z_t \sim i.i.n\left(0, \sigma^2\right) \tag{4.152}$$

which is what we assume with an ARMA(p,q) model. Thus we can view testing H_0 as a diagnostic test of the adequacy of the assumption that the $a_t' s$ are independent.

Engle (1982) has shown that a Lagrange Multiplier (or LM) test of (4.151)is very easy to perform. It consists in running the regression:

$$\hat{a}_t^2 = \beta_0 + \beta_1 \hat{a}_{t-1}^2 + \beta_2 \hat{a}_{t-2}^2 + \cdots + \beta_q \hat{a}_{t-q}^2 + error \tag{4.153}$$

and calculating the R^2 from this regression. Under H_0:

$$LM = T \times R^2 \overset{a}{\sim} \chi_q^2. \tag{4.154}$$

Chapter 5

TS Versus DS Models

5.1 Introduction

We have encountered two methods for dealing with the trend, the trend stationary or TS approach where:

$$TS : X_t = \alpha + \mu t + Y_t \tag{5.1}$$

with $X_t \equiv \ln(W_t)$ (and W_t is say real GNP) and the difference stationary or DS approach where:

$$DS : X_t - X_{t-1} = \mu + Y_t. \tag{5.2}$$

For both the TS and DS models we assume that Y_t is a stationary ARMA(p,q) model: $\phi(B) Y_t = \theta(B) a_t$ with Wold representation:

$$\begin{aligned} Y_t &= \psi(B) a_t \\ &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots . \end{aligned} \tag{5.3}$$

As we have already seen, stationary ARMA(p,q) processes have a number of short-memory properties. Let us quickly review these. The infinite moving average weights ψ_k and the forecast function $E_t[Y_{t+k}]$ have the short-memory property and hence converge to zero very rapidly, essentially exponentially, as k gets large. We write this as $\psi_k = O(\tau^k)$ and $E_t[Y_{t+k}] = O(\tau^k)$ or more precisely:

$$|\psi_k| \leq A\tau^k \tag{5.4}$$

$$|E_t[Y_{t+k}]| \leq C_t \tau^k \tag{5.5}$$

where $0 \leq \tau < 1$ is the absolute value of the largest root of $\phi(B)$. Furthermore long-run uncertainty regarding future Y_t is bounded so that:

$$\begin{aligned} Var_t[Y_{t+k}] &= \sigma^2 \left(1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{k-1}^2\right) \\ &\leq \gamma(0) = \sigma^2 \left(1 + \psi_1^2 + \psi_2^2 + \cdots\right) \end{aligned} \tag{5.6}$$

and $Var_t\left[Y_{t+k}\right] = \gamma(0) + O\left(\tau^{2k}\right)$ so that $Var_t\left[Y_{t+k}\right]$ is bounded by and converges rapidly to $\gamma(0)$.

Since Y_t is modelled the same way for both TS and DS models, there is nothing regarding Y_t that will be nothing qualitatively different between the TS and DS approaches.[1] Rather it is regarding the level of the series: $X_t = \ln(W_t)$ that TS and DS models will have important qualitative differences. In addition we shall see that the asymptotic properties of $\hat{\mu}$, the least squares estimator of μ are very different for TS and DS models.

5.2 Implications of the TS approach

We begin with the implications of adopting the TS approach. We will show that this has the following implications:

1. The effect of past shocks a_{t-k} on X_t is transitory.
2. Forecasts of future X_{t+k} converge rapidly to the deterministic trend $\alpha + \mu(t+k)$.
3. Long-run uncertainty regarding X_{t+k} is bounded.
4. The estimator of μ is superefficient.

5.2.1 Shocks are Transitory

We first show that

Theorem 168 *For the TS model past shocks: a_{t-k} have a transitory impact on X_t; in particular:*

$$\frac{\partial X_t}{\partial a_{t-k}} = \psi_k \ = O\left(\tau^k\right).$$

Proof. From the Wold representation for the TS model in (5.1) we have:

$$\begin{aligned} X_t &= \alpha + \mu t + Y_t & (5.7) \\ &= \alpha + \mu t + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots . \end{aligned}$$

It therefore follows that:

$$\frac{\partial X_t}{\partial a_{t-k}} = \psi_k \ = O\left(\tau^k\right). \qquad (5.8)$$

■

Thus for the TS model $\psi_k \to 0$ exponentially so that past shocks have a transitory impact on X_t

[1] Aside from the fact that we will have different estimates of $\phi(B)$ and $\theta(B)$.

5.2.2 Forecasting

Consider now the problem of forecasting X_{t+k} for the TS model. We have:

Theorem 169 *For the TS model forecasts converge exponentially to the trend line $\alpha + \mu\,(t+k)$. In particular:*

$$E_t[X_{t+k}] = \alpha + \mu(t+k) + O\left(\tau^k\right).$$

Proof. From (5.1) for period $t+k$ we have:

$$E_t[X_{t+k}] = \alpha + \mu(t+k) + E_t[Y_{t+k}]. \tag{5.9}$$

Since $E_t[Y_{t+k}] = O\left(\tau^k\right)$ then result follows. ■

Long-run uncertainty about the future is determined by $Var_t[X_{t+k}]$, which is given by:

$$\begin{aligned} Var_t\left[X_{t+k}\right] &= Var_t\left[\alpha + \mu(t+k) + Y_{t+k}\right] \qquad (5.10)\\ &= Var_t[Y_{t+k}] \\ &= \sigma^2\left(1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{k-1}^2\right). \end{aligned}$$

From this it follows that:

Theorem 170 *If X_t is TS then long-run uncertainty is bounded, in particular:*

$$Var_t\left[X_{t+k}\right] \le \gamma\left(0\right) \tag{5.11}$$

for all forecast horizons k.

Furthermore:

Theorem 171 *$Var_t[X_{t+k}]$ converges rapidly to $\gamma(0)$, in particular:*

$$\left|Var_t[X_{t+k}] - \gamma\left(0\right)\right| = O\left(\tau^{2k}\right).$$

5.2.3 An Example

Suppose that:

$$X_t = 6 + 0.03t + Y_t$$

and Y_t is an AR(1) where

$$Y_t = 0.7Y_{t-1} + a_t.$$

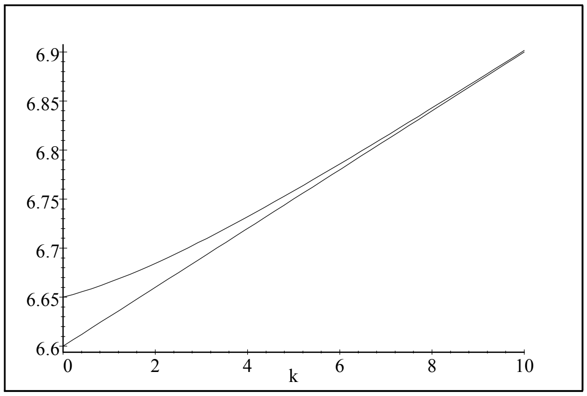
Then

$$E_t\left[Y_{t+k}\right] = (0.7)^k\, Y_t.$$

If $Y_t = 0.05$ then

$$E_t\left[X_{t+k}\right] = 6 + 0.03(t+k) + (0.7)^k\, 0.05$$

which is plotted below using $t = 20$ as the present year.



Forecast and Trend Line

Note how the forecast converges rapidly to the trend line.

Now consider constructing a confidence interval for X_{t+k} based on:

$$E_t\left[X_{t+k}\right] \pm 1.96\sqrt{Var_t\left[X_{t+k}\right]}$$

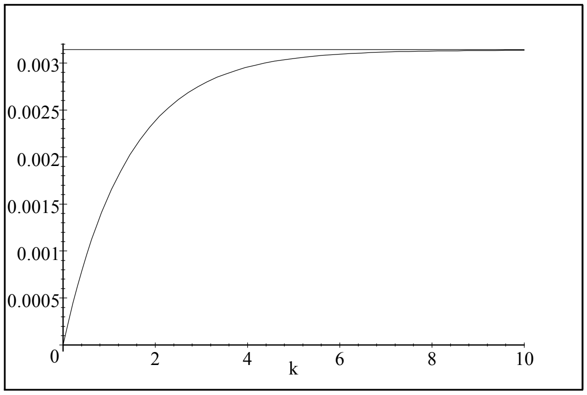
where:

$$Var_t\left[X_{t+k}\right] = (0.04)^2\left(\frac{1-(0.7)^{2k}}{1-(0.7)^2}\right).$$

Note that

$$\gamma(0) = (0.04)^2\left(\frac{1}{1-(0.7)^2}\right) = 0.00314$$

which bounds $Var_t\left[X_{t+k}\right]$ from above as the diagram below illustrates:

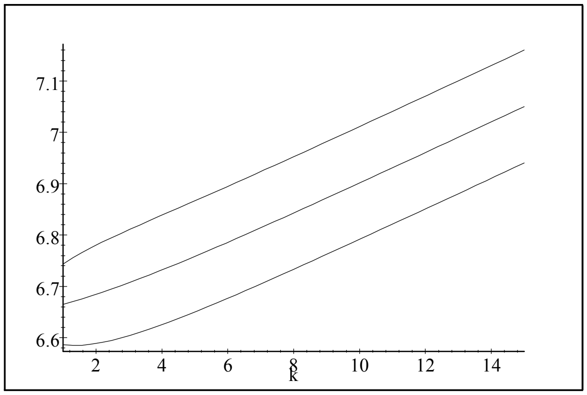


$Var_t\left[X_{t+k}\right]$ and $\gamma(0)$.

Confidence intervals thus take the form:

$$6 + 0.03(t+k) + (0.7)^k\, 0.05 \pm 1.96 \times (0.04)\sqrt{\frac{1-(0.7)^{2k}}{1-(0.7)^2}}.$$

This is plotted below using $t = 20$ as the base period.



Confidence Intervals for X_{t+k}

Notice how the width of the lower and upper bands of the confidence interval quickly reach an upper bound as the forecast horizon k gets larger, reflecting the fact that long-run uncertainty is bounded.

5.2.4 Estimating μ

For the TS the growth rate μ is estimated by regressing X_t on a constant and trend t resulting in the estimator:

$$\hat{\mu} = \frac{\sum_{t=1}^{T}\left(t-\frac{T}{2}\right)\left(X_t - \overline{X}\right)}{\sum_{t=1}^{T}\left(t-\frac{T}{2}\right)^2}. \tag{5.12}$$

It can be shown that ordinary and generalized least squares are asymptotically equivalent for the TS model.

Recall that most reasonable estimators encountered in statistics converge to the population values at a rate of $O_p\left(T^{-\frac{1}{2}}\right)$; that is if $\hat{\beta}$ estimates β then the standard error $SE\left[\hat{\beta}\right] = O\left(T^{-\frac{1}{2}}\right)$ or goes to zero in the same way that as $\frac{1}{\sqrt{T}}$ goes to zero as $T \to \infty$.

Some estimators go to zero at a faster rate than $O_p\left(T^{-\frac{1}{2}}\right)$. These estimators are called superefficient. We have:

Definition 172 *A consistent estimator $\hat{\beta}$ is said to be superefficient if*

$$SE\left[\hat{\beta}\right] = O\left(T^{-\delta}\right)$$

for $\delta > \frac{1}{2}$.

It turns out that one implication of the TS model is that $\hat{\mu}$ is superefficient. In particular we have:

Proposition 173 *If X_t is TS then $\hat{\mu}$ is superefficient, in particular:*

$$SE\left[\hat{\mu}\right] \approx \frac{\sqrt{12}\sigma|\psi(1)|}{T^{\frac{3}{2}}} = O\left(T^{-\frac{3}{2}}\right).$$

Proof. (Informal) It can be shown that:

$$\begin{aligned} Var\left[\hat{\mu}\right] &\approx \frac{\sigma^2|\psi(1)|^2}{\sum_{t=1}^{T}\left(t-\frac{T}{2}\right)^2} &(5.13)\\ &= \frac{\sigma^2|\psi(1)|^2}{\frac{1}{12}T^3+\frac{1}{6}T}\\ &\approx \frac{12\sigma^2|\psi(1)|^2}{T^3}. &(5.14)\end{aligned}$$

The last line follows because the $\frac{1}{12}T^3$ term in the denominator dominates the term $\frac{1}{6}T$ for T large. The asymptotic standard error for $\hat{\mu}$ is therefore given by:

$$SE\left[\hat{\mu}\right] \approx \frac{\sqrt{12}\sigma|\psi(1)|}{T^{\frac{3}{2}}} = O\left(T^{-\frac{3}{2}}\right). \tag{5.15}$$

■

5.3 Implications of the DS approach

We will now show that DS have very different implications. In particular we will show that:

1. The effect of past shocks a_{t-k} on X_t is permanent.
2. Forecasts of future X_{t+k} converge rapidly to a stochastic trend $R_t + \mu k$ where R_t is a random variable.
3. Long-run uncertainty regarding X_{t+k} is unbounded.
4. The Estimator of μ is not superefficient.

5.3.1 The Beveridge-Nelson Decomposition

Proving similar results for the DS model is more difficult because of the presence of X_{t-1} in (5.2) or

$$X_t = \mu + X_{t-1} + Y_t. \tag{5.16}$$

For example a past shock a_{t-k} will affect both X_{t-1} and Y_t which makes it difficult to get the effect of a_{t-k} on X_t by itself.

Of all the DS models random walks are the easiest to analyze. A very useful result is to decompose X_t into a trend and a cycle where the trend is a random walk. This is the Beveridge-Nelson decomposition which states that:

Theorem 174 ***Beveridge-Nelson Decomposition:*** *If X_t is DS as:*

$$(1-B)\,X_t = \mu + \psi\,(B)\,a_t$$

where $\psi_k = O\left(\tau^k\right)$ with $|\tau| < 1$ and $\psi\,(1) \neq 0$ then X_t can be decomposed as:

$$X_t = T_t^* + Y_t^*$$

where T_t^* is a random walk with drift:

$$T_t^* = \mu + T_{t-1}^* + \psi\,(1)\,a_t \tag{5.17}$$

and Y_t^* is a stationary cycle with Wold representation:

$$\begin{aligned} Y_t^* &= \psi^*\,(B)\,a_t \qquad (5.18) \\ &= \psi_o^* a_t + \psi_1^* a_{t-1} + \psi_2^* a_{t-2} + \cdots \end{aligned}$$

and where $\psi_k^* = O\left(\tau^k\right)$.

Proof. See the appendix to this chapter. ∎

Remark 175 *If $\psi\,(1) = 0$ then $T_t^* = \alpha + \mu t$ in which case $X_t = \alpha + \mu t + Y_t^*$ and hence X_t would be TS.*

Remark 176 *The roll played by $\psi\,(1)$ in the Beveridge Nelson decomposition and in many of the results that follow is closely related to the spectrum $f\,(\lambda)$ at frequency $\lambda = 0$ since:*

$$\begin{aligned} f\,(0) &= \left(\frac{\sigma^2}{2\pi}\psi\left(e^{i\lambda}\right)\psi\left(e^{-i\lambda}\right)\right)|_{\lambda=0} \\ &= \frac{\sigma^2}{2\pi}\psi\,(1)^2\,. \end{aligned}$$

(See the section on spectral analysis.) Roughly this means that it is only the low frequency or long-run properties coming from $\psi\,(B)$ that has an affect on the results.

It is possible to calculate the Beveridge-Nelson decomposition for a particular time series in a number of ways. For example:

Theorem 177 *If X_t is DS with Y_t an AR(p) process then T_t^* in the Beveridge-Nelson decomposition is given by:*

$$T_t^* = \frac{1}{\phi\,(1)}\phi\,(B)\,X_t.$$

Proof. From the Beveridge-Nelson decomposition:

$$(1-B)\,T_t^* = \mu + \frac{1}{\phi\,(1)}a_t$$

and since Y_t is an AR(p)

$$\phi(B)(1-B)X_t = \phi(1)\mu + a_t$$

so that:

$$\begin{aligned}(1-B)T_t^* &= \mu + \frac{1}{\phi(1)}\left(\phi(B)(1-B)X_t - \phi(1)\mu\right) \\ &= \frac{\phi(B)}{\phi(1)}(1-B)X_t.\end{aligned}$$

Cancelling the $(1-B)$ from both sides then yields the result. ■

Example 178 *If*

$$X_t = \mu + X_{t-1} + Y_t$$

where: $Y_t = 0.7\ Y_{t-1} + a_t$ then T_t^* can be calculated as:

$$\begin{aligned}T_t^* &= \frac{\phi(B)}{\phi(1)}X_t \\ &= \frac{1-0.7B}{1-0.7}X_t \\ &= \frac{1}{0.3}(X_t - 0.7X_{t-1}).\end{aligned}$$

The cycle Y_t^* can then be calculated as:

$$Y_t^* = X_t - T_t^*.$$

5.3.2 Shocks are Permanent

Unlike TS models past shocks have a permanent impact on X_t. In particular we have:

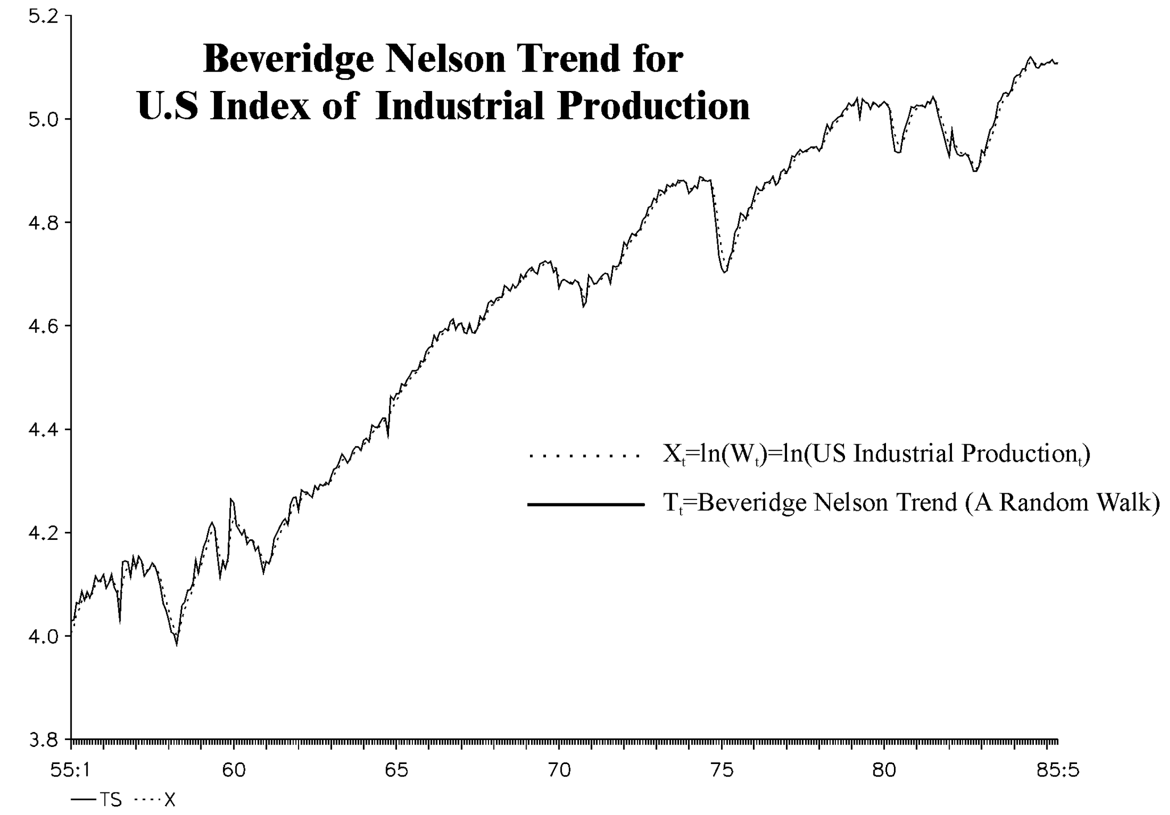
Theorem 179 *If X_t is DS shocks are permanent with*

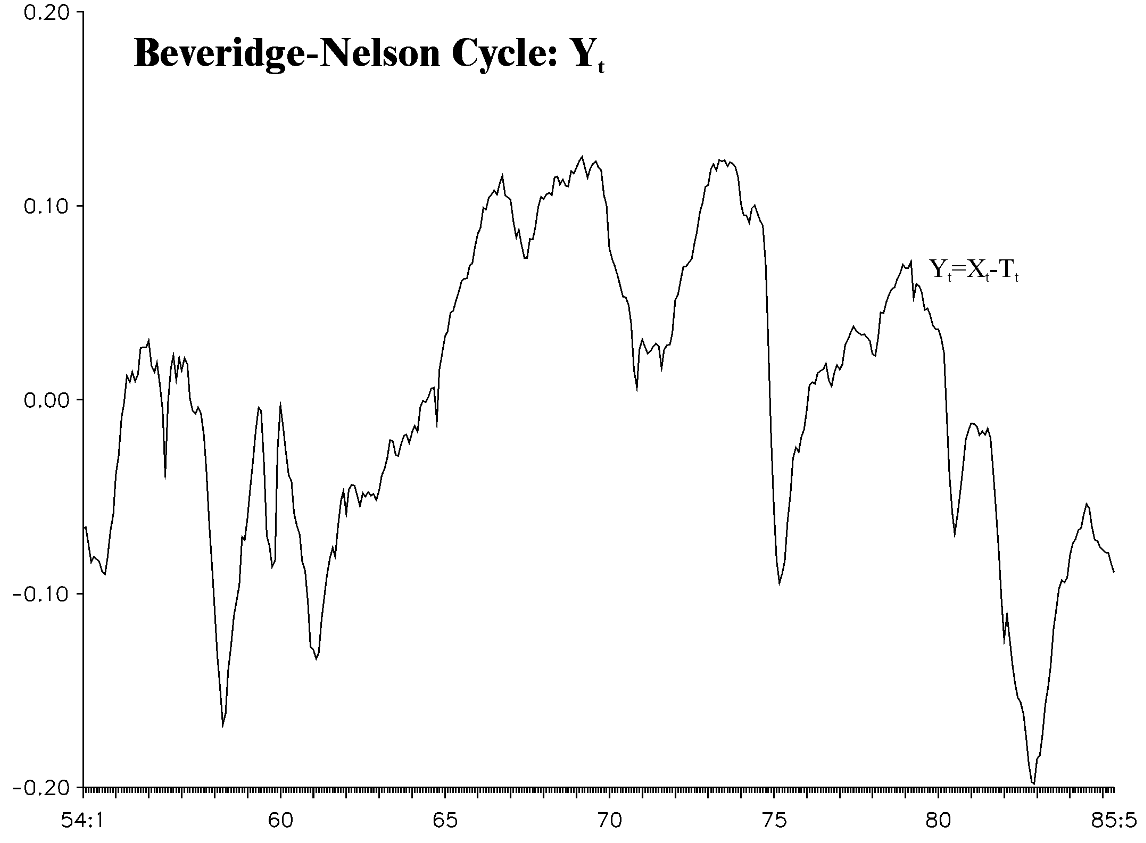
$$\frac{\partial X_t}{\partial a_{t-k}} = \psi(1) + O\left(\tau^k\right) \to \psi(1) \tag{5.19}$$

as $k \to \infty$.

Proof. Since T_t^* in the Beveridge Nelson decomposition is a random walk with drift, it follows that

$$\begin{aligned}T_t^* &= \mu + T_{t-1}^* + \psi(1)a_t \\ &= t\mu + T_o^* + \psi(1)(a_t + a_{t-1} + \cdots + a_1).\end{aligned}$$


Beveridge Nelson Trend for
U.S Index of Industrial Production
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
55:1
60
65
70
75
80
85:5
X_t=ln(W_t)=ln(US Industrial Production$_t$)
T_t=Beveridge Nelson Trend (A Random Walk)
—TS ····X


Beveridge-Nelson Cycle: Y_t
0.20
0.10
0.00
-0.10
-0.20
54:1
60
65
70
75
80
85:5
Y_t=X_t-T_t

Thus a past shock a_{t-k} has a permanent impact on T_t^* with multiplier $\psi(1)$ or:

$$\frac{\partial T_t^*}{\partial a_{t-k}} = \psi(1). \tag{5.20}$$

On the other hand with Y_t^* is stationary and

$$\frac{\partial Y_t^*}{\partial a_{t-k}} = \psi_k^* = O\left(\tau^k\right). \tag{5.21}$$

Since $X_t = T_t^* + Y_t^*$ it follows that:

$$\begin{aligned} \frac{\partial X_t}{\partial a_{t-k}} &= \frac{\partial T_t^*}{\partial a_{t-k}} + \frac{\partial Y_t^*}{\partial a_{t-k}} & (5.22) \\ &= \psi(1) + \psi_k^* \\ &= \psi(1) + O\left(\tau^k\right). & (5.23) \end{aligned}$$

■

Thus there is no tendency for the effect of past shocks on X_t to diminish as we look farther and farther in the past so that shocks have a *permanent* effect on X_t. It is as if the economy today were still affected by a bad harvest 4000 years ago. This is unlike the TS case (where shocks were transitory) shocks

This is one reason why macroeconomists have been so interested in DS models since they imply that shocks such as natural disasters, monetary shocks, demand shocks etc. have permanent effects on the economy; that is their influence does not die out with time.

An Example

Suppose that Y_t follows a stationary AR(1) process:

$$Y_t = \phi Y_{t-1} + a_t \tag{5.24}$$

so that:

$$\psi\left(B\right) = \frac{1}{1 - \phi B}.$$

and hence:

$$\psi\left(1\right) = \frac{1}{1 - \phi}. \tag{5.25}$$

It then follows that:

$$\begin{aligned} \frac{\partial X_t}{\partial a_{t-k}} &= \frac{1 - \phi^k}{1 - \phi} & (5.26) \\ &= \psi\left(1\right) - \frac{\phi^k}{1 - \phi} & (5.27) \\ &= \psi\left(1\right) + O\left(\tau^k\right) & (5.28) \end{aligned}$$

where $\tau = \phi$. Note how $\frac{\partial X_t}{\partial a_{t-k}} \to \psi\left(1\right)$ exponentially as $k \to \infty$.

If $\phi = \frac{1}{2}$ for example then:

$$\psi\left(1\right) = \frac{1}{1-\frac{1}{2}} = 2$$

so that:

$$\begin{aligned} \frac{\partial X_t}{\partial a_{t-k}} &= 2 - 2\left(\frac{1}{2}\right)^k && (5.29) \\ &\to 2. && (5.30) \end{aligned}$$

5.3.3 Forecasting

Now consider the problem of forecasting X_{t+k}. We have:

Theorem 180 *If X_t is DS then forecasts of X_{t+k} converge rapidly to a stochastic trend line or more precisely:*

$$E_t\left[X_{t+k}\right] = T_t^* + \mu k + O\left(\tau^k\right).$$

Thus just as with the TS model, $E_t\left[X_{t+k}\right]$ converges rapidly to a trend line. However, unlike the TS model the trend line is random, in particular the intercept is T_t^* (instead of the constant α for the TS trend $\alpha + \mu t$) which is a nonstationary random walk with drift.

Furthermore for the DS model long-run uncertainty regarding X_{t+k} is unbounded. In particular if we measure uncertainty by the forecast variance then uncertainty increases linearly with the forecast horizon k. In particular:

Theorem 181 *If X_t is DS then:*

$$Var_t[X_{t+k}] = \sigma^2\psi(1)^2 k + O\left(\tau^{2k}\right). \qquad (5.31)$$

Thus for the DS model $Var_t\left[X_{t+k}\right]$ grows linearly with the forecast horizon as $\sigma^2\psi(1)^2 k$.

Proof. Since T_t^* follows a random walk with drift μ we have:

$$T_{t+k}^* = T_t^* + \mu k + \psi\left(1\right)\left(a_{t+k} + a_{t+k-1} + \cdots + a_{t+1}\right) \qquad (5.32)$$

so that:

$$\begin{aligned} E_t\left[T_{t+k}^*\right] &= T_t^* + \mu k && (5.33) \\ Var_t\left[T_{t+k}^*\right] &= \sigma^2\psi\left(1\right)^2 k. \end{aligned}$$

Since Y_t^* is stationary we have:

$$\begin{aligned} E_t\left[Y_{t+k}^*\right] &= O\left(\tau^k\right) \\ Var_t\left[Y_{t+k}^*\right] &= \sigma^2\left(\psi_0^{*2} + \psi_1^{*2} + \psi_2^{*2} + \cdots + \psi_{k-1}^{*2}\right) \\ &\leq Var\left[Y_t^*\right] \end{aligned}$$

and

$$Var_t\left[Y^*_{t+k}\right] = Var\left[Y^*_t\right] + O\left(\tau^{2k}\right).$$

Thus:

$$\begin{aligned} E_t\left[X_{t+k}\right] &= E_t\left[T^*_{t+k}\right] + E\left[Y^*_{t+k}\right] \\ &= T^*_t + \mu k + O\left(\tau^k\right) \end{aligned}$$

and

$$Var_t\left[X_{t+k}\right] = \sigma^2 \psi\left(1\right)^2 k + O\left(\tau^{2k}\right).$$

■

This in turn means that confidence intervals for forecasts for DS models will be approximately:

$$T^*_t + \mu k \pm \sigma|\psi(1)|\sqrt{k} \tag{5.34}$$

so that the width of confidence intervals for X_{t+k} will grow as the square root of the forecast horizon k.

An Example

For example consider the simplest case where Y_t is white noise or $Y_t = a_t$ so that:

$$X_t = X_{t-1} + \mu + Y_t. \tag{5.35}$$

Thus

$$X_t = X_{t-1} + \mu + a_t \tag{5.36}$$

and so X_t is thus a random walk with drift. In the Beveridge Nelson decomposition $X_t = T^*_t$, $Y^*_t = 0$ and $\psi\left(1\right) = 1$.

By repeated substitution we have:

$$X_t = X_0 + \mu t + a_t + a_{t-1} + a_{t-2} + a_{t-3} + \cdots + a_1. \tag{5.37}$$

Consequently we have:

$$\frac{\partial X_t}{\partial a_{t-k}} = 1 \tag{5.38}$$

for all k.

To calculate forecasts note that:

$$X_{t+k} = X_t + \mu k + a_{t+k} + a_{t+k-1} + \cdots + a_{t+1} \tag{5.39}$$

so that the k step ahead forecast is

$$E_t\left[X_{t+k}\right] = X_t + \mu k. \tag{5.40}$$

Note that the intercept for the forecast function is X_t, which is a nonstationary random walk.

Now from (5.39) we have:

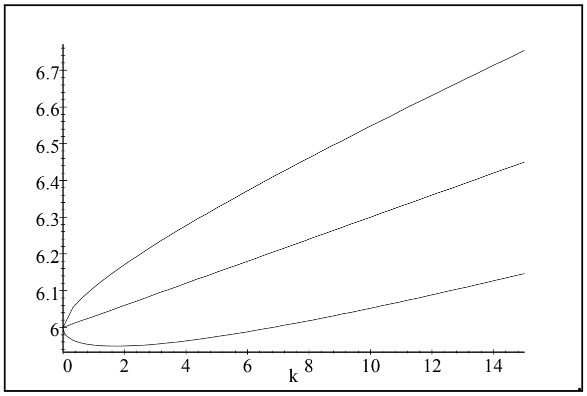
$$Var_t\left[X_{t+k}\right] = \sigma^2 k$$

which then illustrates Theorem 181 since $\psi(1) = 1$.

Suppose then that: $X_t = 6$, $\mu = 0.03$ and $\sigma = 0.04$. A confidence interval for X_{t+k} would take the form:

$$6 + 0.03k \pm 1.96 \times (0.04)\sqrt{k}.$$

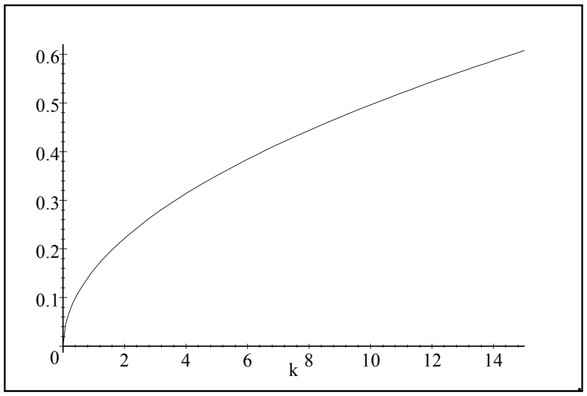
The the forecast and confidence intervals for this DS model are plotted below:



Forecast Confidence Intervals

Note how the width of the confidence interval grows with the square root of the forecast horizon k, reflecting the fact that long-run uncertainty is unbounded.

If we plot the width of the confidence interval against the forecast horizon then this width grows as the square root of the forecast horizon, and so has the shape of a neo-classical production function as illustrated by the diagram below:



Confidence Interval Width

5.3.4 Estimating μ

For DS models the ordinary least squares estimate of μ is given by:

$$\begin{aligned}\hat{\mu} &= \frac{1}{T}\left(\Delta X_1 + \Delta X_2 + \cdots + \Delta X_T\right) \\ &= \frac{X_T - X_0}{T}.\end{aligned}$$

It can be shown that $\hat{\mu}$ is asymptotically equivalent to the generalized least squares estimator of μ.

We have seen that for TS models that $\hat{\mu}$ is superefficient. For DS models $\hat{\mu}$ converges to μ in the conventional $O_p\left(T^{-\frac{1}{2}}\right)$ manner. In particular we have:

Theorem 182 *If X_t is DS then:*

$$SE\left[\hat{\mu}\right] \approx \frac{\sigma|\psi(1)|}{T^{\frac{1}{2}}}.$$

Proof. From the Beveridge-Nelson decomposition for $X_T = T_T^* + Y_T^*$ we have that:

$$\begin{aligned}\hat{\mu} &= \frac{X_T - X_0}{T} \\ &= \frac{T_T^* - T_0^*}{T} + \frac{Y_T^* - Y_0^*}{T}.\end{aligned}$$

Now since:

$$T_T^* = T_o^* + \mu T + \psi(1)\sum_{t=1}^{T} a_t \tag{5.41}$$

it follows that:

$$\hat{\mu} = \mu + \frac{1}{T}\psi(1)\sum_{t=1}^{T} a_t + \frac{Y_T^* - Y_0^*}{T}.$$

Since Y_t^* is stationary it follows that: $\frac{Y_T^* - Y_0^*}{T} \to 0$ in probability so that the variance is dominated by the second term so that:

$$Var\left[\hat{\mu}\right] \approx \frac{\sigma^2 \psi(1)^2}{T}. \tag{5.42}$$

Hence the standard error for $\hat{\mu}$ is:

$$SE\left[\hat{\mu}\right] \approx \frac{\sigma|\psi(1)|}{T^{\frac{1}{2}}}. \tag{5.43}$$

■

5.4 Summary of Differences

We can summarize the difference between TS and DS models then in the following table:

TS	DS
$\frac{\partial X_t}{\partial a_{t-k}} = \psi_k \to 0$ (shocks are transitory)	$\frac{\partial X_t}{\partial a_{t-k}} = \psi(1) + O\left(\tau^k\right) \neq 0$ (shocks are permanent)
$E_t\left[X_{t+k}\right] = \alpha + \mu\left(t+k\right) + O\left(\tau^k\right)$ (intercept is non-random)	$E_t\left[X_{t+k}\right] = T_t^* + \mu k + O\left(\tau^k\right)$ (when $Y_t = a_t$) (intercept is random)
$Var_t\left[X_{t+k}\right] \leq \gamma(0)$ (uncertainty is bounded)	$Var_t\left[X_{t+k}\right] = \sigma^2|\psi(1)|^2 k + O\left(\tau^k\right)$ (uncertainty is unbounded)
$SE\left[\hat{\mu}\right] \approx \frac{\sqrt{12}\sigma|\psi(1)|}{T^{\frac{3}{2}}}$ $\hat{\mu}$ is superefficient	$SE\left[\hat{\mu}\right] \approx \frac{\sigma|\psi(1)|}{T^{\frac{1}{2}}}$ $\hat{\mu}$ is not superefficient

5.5 Testing for Unit Roots

We have seen that many of the properties of the time series X_t depend critically on whether it is TS or DS. Since this is a potentially important issue, and since we often will have no a priori knowledge of whether X_t is TS or DS, it would be nice if we could let the data inform us about what is the appropriate model. In what follows we will construct a regression in which if a particular parameter γ

is zero then X_t is DS while if $\gamma < 0$ X_t is TS. This then leads to the Augmented Dickey-Fuller test (or ADF test) for a unit root.

Imagine that X_t is written as a linear trend plus a cycle Y_t as:

$$X_t = \alpha + \mu t + Y_t \tag{5.44}$$

where we assume that Y_t follows a (not necessarily stationary) AR(p) process:

$$\phi(B) Y_t = a_t. \tag{5.45}$$

Whether X_t is DS or TS will depend on $\phi(B)$. In particular:

Theorem 183 *Suppose that $\phi(B)$ in* (5.44) *is factored as:*

$$\phi(B) = (1 - r_1 B)(1 - r_2 B) \cdots (1 - r_p B)$$

and suppose that $|r_j| < |r_1|$ for $j = 2, 3, \ldots p$ so that r_1 is the largest r_j in absolute value. Then X_t in (5.44) *is DS if and only if $r_1 = 1$ while X_t is TS if and only if $|r_1| < 1$.*

Proof. Clearly $|r_1| < 1$ if and only if Y_t in (5.44) is stationary which in turn is equivalent to X_t being TS. If $r_1 = 1$ then Y_t is not stationary and hence X_t is not TS. However, $(1 - B) Y_t$ is stationary since:

$$\begin{aligned} \phi(B) &= (1 - B)(1 - r_2 B) \cdots (1 - r_p B) \\ &= (1 - B) \tilde{\phi}(B) \end{aligned}$$

where:

$$\tilde{\phi}(B) = (1 - r_2 B)(1 - r_3 B) \cdots (1 - r_p B) \tag{5.46}$$

is a stationary AR(p-1) polynomial. Thus (5.45) becomes:

$$\tilde{\phi}(B)(1 - B) Y_t = a_t$$

and so $(1 - B) Y_t$ is a stationary AR(p-1). Multiplying both sides of (5.44) by $(1 - B)$ we obtain:

$$(1 - B) X_t = \mu + (1 - B) Y_t$$

and so $(1 - B) X_t$ is stationary and hence X_t is DS. To prove the only if part note that if X_t is DS then $(1 - B) Y_t$ must be stationary and hence $r_1 = 1$. ■

Thus to distinguish between TS and DS models we need only to focus on r_1. This however involves calculating roots of polynomials, a very nonlinear computation, and hence is not directly amenable to linear regression methods.

We can link r_1 to a parameter γ defined as follows:

$$\gamma = -\phi(1) = -\left(1 - \phi_1 - \phi_2 - \cdots - \phi_p\right).$$

We now have:

Theorem 184 *Given the assumptions of Theorem 183 then X_t in* (5.44) *is DS if and only if $\gamma = 0$*

Proof. Since:

$$\gamma = -\phi(1) = (1 - r_1)\tilde{\phi}(1)$$

where $\tilde{\phi}(B)$ is defined in (5.46). Since by assumption $\tilde{\phi}(1) \neq 0$ it follows that $\gamma = 0$ if and only if $r_1 = 1$ so by Theorem 183 X_t is DS if and only if $\gamma = 0$. ■

Theorem 185 *Given the assumptions of Theorem 183 then X_t in* (5.44) *is TS if and only if $\gamma < 0$.*

Proof. Since by Theorem 183 X_t is TS if and only if $|r_1| < 1$. Following the proof of Theorem 87 we conclude that this is equivalent to $\phi(1) > 0$ which is equivalent to $\gamma < 0$. ■

Using these two results we can test whether X_t is DS or TS if we can set up a regression where γ is a coefficient on some regressor. This is provided by the next result where γ is the coefficient on X_{t-1} :

Theorem 186 *The series X_t in* (5.44) *can be represented as:*

$$\tilde{\phi}(B)(1 - B)X_t = \delta_o + \delta_1 t + \gamma X_{t-1} + a_t$$

or equivalently as:

$$\Delta X_t = \delta_o + \delta_1 t + \gamma X_{t-1} + \sum_{j=1}^{p-1} \tilde{\phi}_j \Delta X_{t-j} + a_t \tag{5.47}$$

where:

$$\begin{aligned} \delta_o &= \gamma(\alpha - \mu) + \mu\tilde{\phi}(1) \\ \delta_1 &= -\mu\gamma \\ \gamma &= -\phi(1). \end{aligned}$$

To perform the augmented Dickey-Fuller test we run the regression in (5.47), obtain the t statistic on γ, say τ_γ, and test:

$$\begin{aligned} H_o &: X_t \text{ is } DS \text{ (or } \gamma = 0) \text{ versus} \\ H_o &: X_t \text{ is } TS \text{ (or } \gamma < 0) . \end{aligned}$$

If $\tau_\gamma \approx 0$ then this is consistent with X_t being DS while if significantly less than zero or $\tau_\gamma \ll 0$ then this would be consistent with X_t being TS. To perform this test we require a critical value τ_γ^c for our test statistic τ_γ such that if

$$\tau_\gamma > \tau_\gamma^c \tag{5.48}$$

we accept H_o that X_t is DS while if

$$\tau_\gamma < \tau_\gamma^c \tag{5.49}$$

we reject H_o that X_t is DS and instead conclude that X_t is TS.

Because X_t has a unit root under H_o and is therefore not stationary, it turns out that τ_γ does not have an asymptotic standard normal distribution as one would usually expect for a t statistic. Instead τ_γ has a non-normal distribution that is skewed to the left. It is however an easy matter to obtain critical values using computer simulation assuming that one has run the proper regression.

At the 5% significance level the critical value is about:

$$\tau_\gamma^c = -3.4 \tag{5.50}$$

which is much less than the -1.65 one would use if τ_γ were asymptotically standard normal.

Example 187 *If in the regression we obtained a value of $\hat{\gamma} = -0.056$ with t statistic $\tau_\gamma = -1.7$ then since $\tau_\gamma > -3.4$ we would accept H_o that X_t is DS. Alternatively, if we obtained $\hat{\gamma} = -0.25$ with t statistic $\tau_\gamma = -4.7$ then since $\tau_\gamma < -3.4$ we would reject H_o that X_t is DS and conclude that X_t is TS.*

Note that under the null that X_t is DS or that $\gamma = 0$ so that the Dickey-Fuller regression in (5.47) reduces to:

$$\Delta X_t = \mu\tilde{\phi}(1) + \sum_{j=1}^{p-1} \tilde{\phi}_j \Delta X_{t-j} + a_t \tag{5.51}$$

and there is no trend term in the restricted model.

One might then think it is better to run the regression (5.51) instead of (5.47). This would however not be correct since it turns out that the asymptotic distribution of τ_γ would then depend on the unknown μ and so it would not be possible to construct a valid test. It turns out that the only time one should estimate (5.51) instead of (5.47) is if you knew a priori that $\mu = 0$.

In fact all the regressors in (5.51) are needed to insure that τ_γ has the correct asymptotic distribution and does not depend on unknown nuisance parameters. The constant term δ_o is needed to insure that the distribution of τ_γ does not depend on X_o while the lagged values of ΔX_{t-k}, or the "augmented" part of the augmented Dickey-Fuller test, are needed so that the asymptotic distribution of τ_γ does not depend on $\phi(B)$. Roughly speaking one needs a sufficient number of lagged values of ΔX_t to insure that the error term a_t can be considered as being uncorrelated. There are a number of ways of choosing the appropriate number p, one of which would be to use either the Akaike or Schwarz criteria on the DS model .

5.6 Appendix

5.6.1 Proof of Theorem 186

Multiply both sides of

$$X_t = \alpha + \mu t + Y_t$$

by $\phi(B)$ to obtain:

$$\phi(B) X_t = \phi(B)\alpha + \mu\phi(B) t + \phi(B) Y_t.$$

From (5.45)

$$\phi(B) Y_t = a_t.$$

For a constant $Bc = c$ while $Bt = t - 1$ so that:

$$\begin{aligned} \phi(B)\alpha &= \alpha - \phi_1\alpha - \phi_2\alpha - \cdots - \phi_p\alpha && (5.52) \\ &= \phi(1)\alpha \\ \phi(B) t &= t - \sum_{j=1}^{p} \phi_j (t - j) && (5.53) \\ &= \phi(1) t + \sum_{j=1}^{p} \phi_j j \\ &= \phi(1) t - \phi'(1) \end{aligned}$$

where: $\phi'(1)$ is the derivative of $\phi(B)$ evaluated at $B = 1$.

It follows then that (5.44) becomes:

$$\phi(B) X_t = \phi(1)\alpha - \mu\phi'(1) + \mu\phi(1) t + a_t. \tag{5.54}$$

Now if we define $\Gamma(B)$ by:

$$\Gamma(B) = \phi(B) - \phi(1) B \tag{5.55}$$

it is clear that $\Gamma(1) = 0$ so that $(1 - B)$ can be factored out of $\Gamma(B)$ as:

$$\Gamma(B) = \tilde{\phi}(B)(1 - B). \tag{5.56}$$

Since $\phi(0) = 1$ we have $\Gamma(0) = 1$ and hence $\tilde{\phi}(0) = 1$ so that $\tilde{\phi}(B)$ is a $p - 1^{th}$ degree polynomial with $\tilde{\phi}_o = 1$ that can be written as:

$$\tilde{\phi}(B) = 1 - \tilde{\phi}_1 B - \tilde{\phi}_2 B^2 - \cdots - \tilde{\phi}_{p-1} B^{p-1}. \tag{5.57}$$

Then since

$$\Gamma(B) = \phi(B) - \phi(1) B = \tilde{\phi}(B)(1 - B)$$

we can write $\phi(B)$ as:

$$\phi(B) = \phi(1) B + \tilde{\phi}(B)(1-B). \tag{5.58}$$

Finally differentiating $\phi(B)$ at setting $B = 1$ we obtain:

$$\phi'(1) = \phi(1) - \tilde{\phi}(1). \tag{5.59}$$

Inserting (5.58) into (5.54) we obtain:

$$\left(\phi(1) B + \tilde{\phi}(B)(1-B)\right) X_t = \phi(1)\alpha + \mu\left(\tilde{\phi}(1) - \phi(1)\right) + \mu\phi(1) t + a_t \tag{5.60}$$

or equivalently:

$$\tilde{\phi}(B)(1-B) X_t = \phi(1)\alpha + \mu\left(\tilde{\phi}(1) - \phi(1)\right) + \mu\phi(1) t + a_t. \tag{5.61}$$

5.6.2 Proof of the Beveridge Nelson Decomposition

Let $(1-B)X_t$ be a stationary time series with Wold representation:

$$(1-B) X_t = \mu + \psi(B) a_t. \tag{5.62}$$

where we assume that:

$$|\psi_k| < A\tau^k \tag{5.63}$$

for some A and τ such that

$$0 \leq \tau < 1. \tag{5.64}$$

In addition we assume that:

$$\psi(1) \neq 0. \tag{5.65}$$

Together these assumptions imply that X_t is DS.

We can rewrite the model as:

$$\begin{aligned} (1-B) X_t &= \mu + \psi(1) a_t + (\psi(B) - \psi(1)) a_t \\ &= (1-B) T_t^* + (1-B) Y_t^* \end{aligned} \tag{5.66}$$

where we define the trend T_t^* by:

$$(1-B) T_t^* = \mu + \psi(1) a_t \tag{5.67}$$

and the cycle Y_t^* by:

$$Y_t^* = \psi^*(B) a_t \tag{5.68}$$

where:

$$\psi^*(B) = \frac{(\psi(B) - \psi(1))}{1 - B}. \tag{5.69}$$

Cancelling $(1 - B)$ from both sides of (5.66) yields:

$$X_t = T_t^* + Y_t^* \tag{5.70}$$

which is the Beveridge-Nelson decomposition. What remains to be shown is that T_t^* is a random walk with drift while Y_t^* is stationary.

It is easy to show that T_t^* follows a random walk with drift μ since from it's definition we have:

$$T_t^* = \mu + T_{t-1}^* + \psi(1)\, a_t. \tag{5.71}$$

To show that Y_t^* is stationary note that from (5.69) that:

$$(1 - B)\,\psi^*(B) = \psi(B) - \psi(1) \tag{5.72}$$

so that equating powers of B^k on both sides of:

$$(1 - B)\left(\sum_{j=0}^{\infty} \psi_j^* B^j\right) = \sum_{j=0}^{\infty} \psi_j B^j - \psi(1) \tag{5.73}$$

we obtain:

$$\begin{aligned} \psi_k^* - \psi_{k-1}^* &= \psi_k \\ \psi_0^* &= 1 - \psi(1) = -\sum_{j=1}^{\infty} \psi_j. \end{aligned} \tag{5.74}$$

From this we conclude that:

$$\psi_k^* = -\sum_{j=k+1}^{\infty} \psi_j. \tag{5.75}$$

Now using the assumption that $\psi_k = O\left(\tau^k\right)$ or $|\psi_k| \leq A\tau^k$ or has the short-memory property, we show $\psi_k^* = O\left(\tau^k\right)$ or $|\psi_k^*| \leq D\tau^k$ or ψ_k^* also has the short-memory property.

We have:

$$\begin{aligned} |\psi_k^*| &= \left| \sum_{j=k+1}^{\infty} \psi_j \right| \qquad (5.76) \\ &\leq \sum_{j=k+1}^{\infty} A\tau^j \\ &= A \sum_{j=k+1}^{\infty} \tau^j \\ &= \frac{A\tau^{k+1}}{1-\tau} \\ &= D\tau^k \end{aligned}$$

where:

$$D = \frac{A\tau}{1-\tau}. \qquad (5.77)$$

We therefore conclude from Theorem 58 that:

$$Y_t^* = \psi^* \left(B \right) a_t \qquad (5.78)$$

is a stationary time series with a finite variance. This completes the proof of the Beveridge-Nelson decomposition.

Chapter 6

Multivariate Time Series

6.1 VARMA(p,q) Models

6.1.1 Introduction

In economics we are often interested in the relationship between different time series. For example we might be interested in the relationship between Y_{1t} or national income: GNP_t, Y_{2t} the money supply M_t^s and Y_{3t} the rate of interest R_t (all variables detrended say using either TS or DS detrending).

Consider then a straightforward generalization of a scalar ARMA(p,q). Let Y_t be an $n \times 1$ vector of time series and a_t be an $n \times 1$ vector of *i.i.d.* shocks so that:

$$Y_t = \begin{bmatrix} Y_{1t} \\ Y_{2t} \\ \vdots \\ Y_{nt} \end{bmatrix}, \ a_t = \begin{bmatrix} a_{1t} \\ a_{2t} \\ \vdots \\ a_{nt} \end{bmatrix}.$$

For example if we are interested in GNP, the money supply and interest rates then $n = 3$ and

$$Y_t = \begin{bmatrix} Y_{1t} \\ Y_{2t} \\ Y_{3t} \end{bmatrix} = \begin{bmatrix} GNP_t \\ M_t^s \\ R_t \end{bmatrix}, \ a_t = \begin{bmatrix} a_{1t} \\ a_{2t} \\ a_{3t} \end{bmatrix}$$

so that a_{1t} would then be the GNP shock, a_{2t} the money shock and a_{3t} the interest rate shock.

We can then have a vector ARMA(p,q) process or VARMA(p,q) defined as:

Definition 188 ***VARMA(p,q)****: Y_t follows a VARMA(p,q) process if:*

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=1}^{q} \theta_j a_{t-j} + a_t \tag{6.1}$$

where a_t is an $n \times 1$ vector of i.i.d. shocks with:

$$a_t \sim i.i.n(0, \Omega)$$

where Ω is $n \times n$ symmetric and positive definite while ϕ_j and θ_j are $n \times n$ matrices.

The density of a_t is given by:

$$p\left(a_t\right) = (2\pi)^{-\frac{n}{2}} \left|\Omega\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} a_t^T \Omega^{-1} a_t\right). \tag{6.2}$$

where the inverse of Ω exists since Ω is positive definite. We can write Ω as:[1]

$$\underset{n\times n}{\Omega} = \begin{bmatrix} \omega_{11} & \omega_{12} & \dots & \omega_{1n} \\ \omega_{12} & \omega_{22} & \dots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{1n} & \omega_{2n} & \dots & \omega_{nn} \end{bmatrix} \tag{6.3}$$

so that:

$$\begin{aligned} Var\left[a_{it}\right] &= \omega_{ii} \\ Cov\left[a_{it}, a_{jt}\right] &= \omega_{ij}. \end{aligned} \tag{6.4}$$

Thus ω_{ii} is the variance of the i^{th} shock while ω_{ij} is the contemporaneous covariance between the i^{th} and the j^{th} shock. Thus although a_t is independent across time, we allow for contemporaneous correlation between the shocks at time t.

We can write the VARMA(p,q) model more compactly as:

$$\phi\left(B\right) Y_t = \theta\left(B\right) a_t \tag{6.5}$$

where $\phi\left(B\right)$ and $\theta\left(B\right)$ are matrix polynomials given by:

$$\begin{aligned} \phi\left(B\right) &= I - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \\ \theta\left(B\right) &= I + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q. \end{aligned} \tag{6.6}$$

Note that B can be treated as a scalar when manipulating $\phi\left(B\right)$ and $\theta\left(B\right)$.

The (k, l) elements of the square matrices ϕ_j and θ_j will be denoted by ϕ_{kl}^j and θ_{kl}^j so that:

$$\phi_j = \begin{bmatrix} \phi_{11}^j & \phi_{12}^j & \dots & \phi_{1n}^j \\ \phi_{21}^j & \phi_{22}^j & \dots & \phi_{2n}^j \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1}^j & \phi_{n2}^j & \dots & \phi_{nn}^j \end{bmatrix}, \; \theta_j = \begin{bmatrix} \theta_{11}^j & \theta_{12}^j & \dots & \theta_{1n}^j \\ \theta_{21}^j & \theta_{22}^j & \dots & \theta_{2n}^j \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{n1}^j & \theta_{n2}^j & \dots & \theta_{nn}^j \end{bmatrix}. \tag{6.7}$$

We can also define the vector analogues of AR(p)'s and MA(q)'s as:

[1] Note that for the univariate case $(n = 1)$ we had: $\Omega = \sigma^2$. It might therefore be natural to use Σ rather than Ω . We use Ω rather than Σ to avoid a conflict with the summation notation $\sum_{i=1}^n$.

Definition 189 ***VAR(p):*** *Y_t follows a VAR(p) process if:*

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + a_t. \tag{6.8}$$

or $\phi(B) Y_t = a_t$.

Definition 190 ***VMA(q):*** *Y_t follows a VMA(q) process if:*

$$Y_t = \sum_{j=1}^{q} \theta_j a_{t-j} + a_t$$

or $Y_t = \theta(B) a_t$.

An Example

For example if $n = 2$ then Y_{1t} might be GNP while Y_{2t} might be money (appropriately detrended by either DS or TS). Then a VARMA$(2,1)$ with $n = 2$ components in Y_t would be:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + a_t + \theta_1 a_{t-1} \tag{6.9}$$

or unpacking the matrix notation:

$$\begin{aligned} \begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} &= \begin{bmatrix} \phi_{11}^1 & \phi_{12}^1 \\ \phi_{21}^1 & \phi_{22}^1 \end{bmatrix} \begin{bmatrix} Y_{1t-1} \\ Y_{2t-1} \end{bmatrix} + \begin{bmatrix} \phi_{11}^2 & \phi_{12}^2 \\ \phi_{21}^2 & \phi_{22}^2 \end{bmatrix} \begin{bmatrix} Y_{1t-2} \\ Y_{2t-2} \end{bmatrix} \qquad (6.10) \\ &+ \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} + \begin{bmatrix} \theta_{11}^1 & \theta_{12}^1 \\ \theta_{21}^1 & \theta_{22}^1 \end{bmatrix} \begin{bmatrix} a_{1t-1} \\ a_{2t-1} \end{bmatrix} \end{aligned}$$

or multiplying out the matrices:

$$\begin{aligned} Y_{1t} &= \phi_{11}^1 Y_{1t-1} + \phi_{12}^1 Y_{2t-1} + \phi_{11}^2 Y_{1t-2} + \phi_{12}^2 Y_{2t-2} + a_{1t} + \theta_{11}^1 a_{1t-1} + \theta_{12}^1 a_{2t-1} \\ Y_{1t} &= \phi_{21}^1 Y_{1t-1} + \phi_{22}^1 Y_{2t-1} + \phi_{21}^2 Y_{1t-2} + \phi_{22}^2 Y_{2t-2} + a_{2t} + \theta_{21}^1 a_{1t-1} + \theta_{22}^1 a_{2t-1}. \end{aligned}$$

We then have lagged GNP and money up to lag 2 in both the GNP and the money equations as well as lagged GNP shocks a_{1t-1} and money shocks a_{2t-1} in both equations. Furthermore if $\omega_{12} \neq 0$ there will be a contemporaneous correlation between the GNP shock a_{1t} and the money shock a_{2t}.

Using the notation in (6.6) we can write this VARMA(2,1) as: $\phi(B) Y_t = \theta(B) a_t$ where:

$$\begin{aligned} \phi(B) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \phi_{11}^1 & \phi_{12}^1 \\ \phi_{21}^1 & \phi_{22}^1 \end{bmatrix} B - \begin{bmatrix} \phi_{11}^2 & \phi_{12}^2 \\ \phi_{21}^2 & \phi_{22}^2 \end{bmatrix} B^2 \qquad (6.11) \\ \theta(B) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \theta_{11}^1 & \theta_{12}^1 \\ \theta_{21}^1 & \theta_{22}^1 \end{bmatrix} B. \end{aligned}$$

6.1.2 Stationarity

We now investigate the conditions which determine whether a $VARMA(p,q)$ process is stationary. As you might expect stationarity depends on $\phi(B)$. There is a complication however in that $\phi(B)$ is a matrix polynomial and not a scalar polynomial. The appropriate generalization involves the roots of the determinant of $\phi(B)$ defined as:

Definition 191 *Define the scalar polynomial $\alpha(B)$ as:*

$$\alpha(B) = \det[\phi(B)] \tag{6.12}$$

which is a scalar polynomial of order $p^* = np$ so that:

$$\alpha(B) = 1 + \alpha_1 B + \alpha_2 B^2 + \cdots + \alpha_{p^*} B^{p^*}$$

where α_i is a scalar.

We can always rewrite a VARMA(p,q) process so that it has a scalar autoregressive polynomial $\alpha(B)$ as:

Theorem 192 *A $VARMA(p,q)$ process: $\phi(B)Y_t = \theta(B)a_t$ can always be written as:*

$$\alpha(B)Y_t = \tilde{\theta}(B)a_t \tag{6.13}$$

where $\alpha(B)$ is given in (6.12) and $\tilde{\theta}(B)a_t$ is a finite order vector moving average process of order $q + p(n-1)$.

Proof. For any square matrix A

$$adj[A]\ \ A = A\, adj[A] = \det[A]\, I \tag{6.14}$$

where $adj[A]$ is the adjoint matrix of A. Now if we multiply both sides of $\phi(B)Y_t = \theta(B)a_t$ with $adj[\phi(B)]$ we obtain:

$$\alpha(B)Y_t = adj[\phi(B)]\,\theta(B)a_t. \tag{6.15}$$

Since $adj[\phi(B)]$ involves taking determinants of sub-matrices of $\phi(B)$ of order $(n-1)\times(n-1)$, it is a finite order matrix polynomial which can be written as:

$$adj[\phi(B)] = \sum_{j=0}^{q^*} \tilde{\theta}_j B^j. \tag{6.16}$$

and where $q^* = p(n-1)$. Therefore

$$\tilde{\theta}(B) = adj[\phi(B)]\,\theta(B)$$

is a finite order matrix polynomial of order $q + q^* = q + p(n-1)$. ■

An implication of this result is that

Corollary 193 *If Y_{it} is an element of a vector VARMA(p,q) process then Y_{it} is a scalar ARMA(pn,q+p(n-1)).*

Example 194 *If Y_t is an VARMA(3,1) and Y_{it} is an element of Y_t with $n = 10$ time series in Y_t, then Y_{it} will be an ARMA(30,28) model since $pn = 30$ and $q + p(n-1) = 28$.*

Since the right-hand side of (6.13) is a finite order vector moving average, it is always stationary. We therefore have:

Theorem 195 *A necessary condition for a $VARMA(p,q)$ process to be stationary is that the scalar polynomial $\alpha(B)$ have roots all greater than 1 in absolute value or:*

$$\det[\phi(B)] = 0 \Longrightarrow |B| > 1.$$

6.1.3 An Example

To see how this works consider a $VAR(1)$ with $n = 2$:

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} Y_{1t-1} \\ Y_{2t-1} \end{bmatrix}.$$

Since:

$$\begin{aligned} \phi(B) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} B \\ &= \begin{bmatrix} 1-\phi_{11}B & -\phi_{12}B \\ -\phi_{21}B & 1-\phi_{22}B \end{bmatrix} \end{aligned} \tag{6.17}$$

we have:

$$\begin{aligned} \tilde{\theta}(B) &= adj[\phi(B)] \\ &= \begin{bmatrix} 1-\phi_{22}B & \phi_{12}B \\ \phi_{21}B & 1-\phi_{11}B \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \underbrace{\begin{bmatrix} -\phi_{22} & \phi_{12} \\ \phi_{21} & -\phi_{11} \end{bmatrix}}_{\tilde{\theta}_1} B \end{aligned} \tag{6.18}$$

so that $q^* = 1$.

We can also find the scalar polynomial $\alpha(B)$ as:

$$\begin{aligned} \alpha(B) &= \det[\phi(B)] \\ &= \det \begin{bmatrix} 1-\phi_{11}B & -\phi_{12}B \\ -\phi_{21}B & 1-\phi_{22}B \end{bmatrix} \\ &= (1-\phi_{11}B)(1-\phi_{22}B) - \phi_{12}\phi_{21}B^2 \\ &= 1 - \underbrace{(\phi_{11}+\phi_{22})}_{\alpha_1}B - \underbrace{(\phi_{12}\phi_{21} - \phi_{11}\phi_{22})}_{\alpha_2}B^2 \end{aligned}$$

so that $p^* = 2$.

We can thus rewrite the $VAR(1)$ model as:

$$\alpha(B) Y_t = \tilde{\theta}(B) a_t \tag{6.19}$$

or:

$$\alpha(B) \begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \phi_{22} & \phi_{12} \\ \phi_{21} & \phi_{11} \end{bmatrix} B \right) \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} \tag{6.20}$$

or as:

$$\begin{aligned} \alpha(B) Y_{1t} &= a_{1t} + \phi_{22} a_{1t-1} + \phi_{12} a_{2t-1} \\ \alpha(B) Y_{2t} &= a_{2t} + \phi_{21} a_{1t-1} + \phi_{11} a_{2t-1}. \end{aligned}$$

The right-hand side of both equations is a scalar MA(1) so that both Y_{1t} and Y_{2t} follow scalar ARMA(2,1) processes.

If then we have:

$$\begin{aligned} Y_{1t} &= 0.5 Y_{1t-1} + 0.1 Y_{2t-1} + a_{1t} \\ Y_{2t} &= 0.2 Y_{1t-1} + 0.3 Y_{2t-1} + a_{2t} \end{aligned} \tag{6.21}$$

then:

$$\begin{aligned} \alpha(B) &= \det[\phi(B)] \\ &= \det \begin{bmatrix} 1 - 0.5B & -0.1B \\ -0.2B & 1 - 0.3B \end{bmatrix} \\ &= 1 - 0.8B + 0.13B^2 \end{aligned}$$

and so $p^* = 2$ here. The two roots of $\alpha(B)$ are given by:

$$\begin{aligned} B_1 &= \frac{0.8 + \sqrt{(-0.8)^2 - 4(0.13)}}{2(0.13)} = 4.41 \\ B_2 &= \frac{0.8 - \sqrt{(-0.8)^2 - 4(0.13)}}{2(0.13)} = 1.74. \end{aligned} \tag{6.22}$$

Since $|B_1| = 4.41 > 1$ and $|B_2| = 1.74 > 1$ this process is stationary.

6.1.4 Wold Representation

As with univariate time series we have the Wold representation:

Theorem 196 ***Wold Representation:*** *Every stationary vector time series Y_t with $E[Y_t] = 0$ has an infinite moving average representation:*

$$\begin{aligned} Y_t &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \psi_3 a_{t-3} + \cdots \\ &= \psi(B) a_t \end{aligned} \tag{6.23}$$

where a_t is an uncorrelated series with $E[a_t] = 0$,

$$Var[a_t] = E\left[a_t a_t^T\right] = \Omega$$

and:

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 + \cdots .$$

We can calculate the Wold representation for a VAR(p) just as with an AR(p) as follows:

Theorem 197 ***Recursive Calculation of*** ψ_k*: For a stationary VAR(p) process $\phi(B) Y_t = a_t$ the Wold representation can be recursively calculated as:*

$$\psi_k = \phi_1 \psi_{k-1} + \phi_2 \psi_{k-2} + \cdots + \phi_p \psi_{k-p}$$

with starting values:

$$\psi_0 = I, \ \psi_k = 0 \textit{ for } k < 0.$$

It can then be shown that:

Theorem 198 *For an VAR(p) the infinite moving average weights given by the $n \times n$ matrix: ψ_k can be expressed as:*

$$\psi_k = A_1 r_1^k + A_2 r_2^k + \cdots + A_{p^*} r_{p^*}^k$$

where: r_i^{-1} is one of the p^* roots of $\alpha(B) = \det[\phi(B)]$ and A_i are $n \times n$ matrices.

6.1.5 VAR(1) as a General Special Case

It is easy to derive the Wold representation for the VAR(1) model as

$$\begin{aligned} Y_t &= \phi Y_{t-1} + a_t \\ &= a_t + \phi a_{t-1} + \phi^2 Y_{t-2} \end{aligned}$$

and continuing this process we have:

$$Y_t = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \phi^3 a_{t-3} + \cdots$$

so that:

Theorem 199 *For an VAR(1) process:*

$$\psi_k = \phi^k.$$

Stationarity requires that $\psi_k \to 0$ as $k \to \infty$. If we write ϕ as:

$$\phi = C \Lambda C^{-1}$$

where Λ is a diagonal matrix with the eigenvalues of ϕ along the diagonal then it is easily verified that:

$$\phi^k = C \Lambda^k C^{-1}.$$

Therefore $\phi^k \to 0$ is equivalent to $\Lambda^k \to 0$ and so we have:

Theorem 200 *A VAR(1) process is stationary if and only if all the eigenvalues of ϕ are less than* 1 *in absolute value.*

This result actually provides us with an alternative criterion for stationarity for *all* VAR(p) processes (and indeed of any VARMA(p,q) process). This is because of the following result:

Theorem 201 *Any VARMA(p) process can be represented by an VAR(1) process.*

Proof. We will only prove this for a VAR(p) process. Suppose Y_t follows a VAR(p) process so that:

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + a_t.$$

Define $\tilde{Y}_t$, $\tilde{\phi}$ and $\tilde{a}_t$ as:

$$\tilde{Y}_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p+1} \end{bmatrix}, \tilde{\phi} = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_p \\ I & 0 & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & I & 0 \end{bmatrix}, \tilde{a}_t = \begin{bmatrix} a_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

then the VAR(p) can be written as:

$$\begin{bmatrix} Y_t \\ Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_p \\ I & 0 & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & I & 0 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ Y_{t-3} \\ \vdots \\ Y_{t-p} \end{bmatrix} + \begin{bmatrix} a_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

or as $\tilde{Y}_t = \tilde{\phi}\tilde{Y}_{t-1} + \tilde{a}_t$. ■

Remark 202 *The VAR(1) representation of a time series model is called the* state space representation*. The state space representation in turn is the basis for the* Kalman filter*, a very powerful algorithm which can be used, for example, to obtain the* exact *likelihood of any VARMA(p,q) process and hence to obtain estimates, to handle measurement error and data with different timing intervals (for example monthly and quarterly data). See Harvey's book* Forecasting, Structural Time Series Models and the Kalman Filter *for a good introduction to this topic.*

Since we have a stationarity condition for a VAR(1) in Theorem 200 and since any VAR(p) can be written as a VAR(1) we have:

Theorem 203 *A VAR(p) process is stationary only if all the eigenvalues of $\tilde{\phi}$ are less than* 1 *in absolute value.*

Theorem 204 *The eigenvalues of $\tilde{\phi}$ are the inverse of the roots of $\alpha(B)$. In particular if λ is an eigenvalue of $\tilde{\phi}$ if and only if:*

$$\alpha\left(\lambda^{-1}\right) = 0.$$

6.1.6 Some Examples

Example 1

Consider a scalar AR(2) process:

$$Y_t = 0.6Y_{t-1} + 0.2Y_{t-2} + a_t. \tag{6.24}$$

This can be written as a VAR(1) as:

$$\begin{bmatrix} Y_t \\ Y_{t-1} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \end{bmatrix} + \begin{bmatrix} a_t \\ 0 \end{bmatrix}.$$

We can check for stationarity by calculating the eigenvalues of

$$\tilde{\phi} = \begin{bmatrix} 0.6 & 0.2 \\ 1 & 0 \end{bmatrix}$$

or by finding the roots of

$$\det\left[\begin{bmatrix} 0.6 & 0.2 \\ 1 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right] = 0.$$

Since these eigenvalues are $\lambda_1 = 0.838$ and $\lambda_2 = -0.238$ and both are less than 1 in absolute value, we conclude that the process is stationary.

Example 2

Suppose now that we have $VAR(2)$

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + a_t$$

with $n = 2$ and suppose further that: ϕ_1 and ϕ_2 are given by:

$$\phi_1 = \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}, \ \phi_2 = \begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}.$$

First let us check for stationarity by calculating the roots of $\alpha(B)$ given by:

$$\begin{aligned} \alpha(B) &= \det\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - B\begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix} - B^2\begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}\right) \\ &= 1 - 0.8B - 0.2B^2 + .02B^3 + .01B^4. \end{aligned}$$

Using the computer we find that the four roots of $\alpha(B)$ are:

$$\begin{aligned} B_1 &= -3.9018 + 2.2711i,\ B_2 = -3.9018 - 2.2711i \\ B_3 &= 1.0272,\ B_4 = 4.7763. \end{aligned}$$

Since all roots are greater than 1 in absolute value[2] we conclude that the estimated model is stationary.

Alternatively we can check for stationarity by finding the eigenvalues of the matrix:

$$\tilde{\phi} = \begin{bmatrix} \phi_1 & \phi_2 \\ I & 0 \end{bmatrix}$$

which here is given by:

$$\tilde{\phi} = \begin{bmatrix} 0.5 & -0.1 & 0.2 & 0.3 \\ 0.5 & 0.3 & 0.1 & 0.2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Using the computer we find the eigenvalues are:

$$\begin{aligned} \lambda_1 &= -0.19143 + 0.11143i,\ \lambda_2 = -0.19143 - 0.11143i \\ \lambda_3 &= 0.9735,\ \lambda_4 = 0.20937. \end{aligned}$$

Since all eigenvalues are less than 1 in absolute value we conclude the process is stationary.

6.2 VAR(p) Estimation

6.2.1 Linear Regression of VAR(p) Models

In general estimating and identifying vector VARMA(p,q) with a moving average component ($q > 0$) processes is quite difficult. Most applied work deals with the VAR(p) model:

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + a_t \tag{6.25}$$

which can be written in scalar notation as:

$$Y_{it} = \sum_{k=1}^{p} \sum_{j=1}^{n} \phi_{ij}^{k} Y_{jt-k} + a_{it}\ ,\ \ i = 1, 2, \ldots n. \tag{6.26}$$

[2] Note that for the two complex roots that:

$$|B_1| = |B_2| = |-3.901\,8 \pm 2.\,271\,1i| = 4.5146 > 1$$

Just as with an AR(p), a VAR(p) can be estimated by ordinary least squares where the lagged series Y_{jt-k} act as regressors.

For example with $n = 2$ variables (say GNP and money appropriately detrended) and $p = 2$ lags we have:

$$\begin{aligned}\text{GNP} \quad &: \quad Y_{1t} = \underbrace{\phi_{11}^1 Y_{1t-1} + \phi_{12}^1 Y_{2t-1}}_{\text{Lag 1 GNP and Money}} + \underbrace{\phi_{11}^2 Y_{1t-2} + \phi_{12}^2 Y_{2t-2}}_{\text{Lag 2 GNP and Money}} + a_{1t} \quad (6.27)\\ \text{Money} \quad &: \quad Y_{2t} = \underbrace{\phi_{21}^1 Y_{1t-1} + \phi_{22}^1 Y_{2t-1}}_{\text{Lag 1 GNP and Money}} + \underbrace{\phi_{21}^2 Y_{1t-2} + \phi_{22}^2 Y_{2t-2}}_{\text{Lag 2 GNP and Money}} + a_{2t}.\end{aligned}$$

Instead of grouping the variables on the right-hand side by lag, we could group them by variable (e.g. GNP and Money). This correspond to reversing the double summation and writing:

$$Y_{it} = \sum_{j=1}^{n} \sum_{k=1}^{p} \phi_{ij}^k Y_{jt-k} + a_{it} \quad i = 1, 2, \ldots n. \tag{6.28}$$

In the above example this would mean that we write:

$$\begin{aligned}\text{GNP} \quad &: \quad Y_{1t} = \underbrace{\phi_{11}^1 Y_{1t-1} + \phi_{11}^2 Y_{1t-2}}_{\text{GNP: all lags}} + \underbrace{\phi_{12}^1 Y_{2t-1} + \phi_{12}^2 Y_{2t-2}}_{\text{Money: all lags}} + a_{1t} \quad (6.29)\\ \text{Money} \quad &: \quad Y_{2t} = \underbrace{\phi_{21}^1 Y_{1t-1} + \phi_{21}^2 Y_{1t-2}}_{\text{GNP: all lags}} + \underbrace{\phi_{22}^1 Y_{2t-1} + \phi_{22}^2 Y_{2t-2}}_{\text{Money: all lags}} + a_{2t}.\end{aligned}$$

Since a_{1t} and a_{2t} are uncorrelated with lagged GNP and lagged money, we can estimate this model by running ordinary least squares on the GNP equation and on the money equation. As long as the VAR is stationary this will lead to consistent estimates with the usual asymptotic properties. The complication is that the error term in the GNP equation a_{1t} may be contemporaneously correlated (with a covariance ω_{12}) with the error term in the money equation a_{2t}. The question is could we obtain more efficient estimates if we estimated both the GNP and money equation together as a system.

Before answering this question let us first treat the general case where instead of 2 variables we have n variables in the VAR. Taking the transpose of both sides of (6.25) we obtain:

$$\begin{aligned} Y_t^T &= \sum_{k=1}^{p} Y_{t-k}^T \phi_k^T + a_t^T \qquad (6.30)\\ &= \begin{bmatrix} Y_{t-1}^T \; Y_{t-2}^T \; Y_{t-3}^T \; \cdots Y_{t-p}^T \end{bmatrix} \begin{bmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_p^T \end{bmatrix} + a_t^T \\ &= X_t^T \phi + a_t \end{aligned}$$

where

$$\underset{np\times 1}{X_t} = \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ Y_{t-3} \\ \vdots \\ Y_{t-p} \end{bmatrix} \quad \text{and} \quad \underset{np\times p}{\phi} = \begin{bmatrix} \phi_1^T \\ \phi_2^T \\ \phi_3^T \\ \vdots \\ \phi_p^T \end{bmatrix}. \tag{6.31}$$

We can think of X_t as a vector of regressors and ϕ as a matrix of coefficients to estimate. In the example in (6.27) we would have:

$$X_t = \begin{bmatrix} Y_{1t-1} \\ Y_{2t-1} \\ Y_{1t-2} \\ Y_{2t-2} \end{bmatrix}, \ \phi = \begin{bmatrix} \phi_{11}^1 & \phi_{21}^1 \\ \phi_{12}^1 & \phi_{22}^1 \\ \phi_{11}^2 & \phi_{21}^2 \\ \phi_{12}^2 & \phi_{22}^2 \end{bmatrix}. \tag{6.32}$$

If we now define:

$$\underset{T\times n}{Y} = \begin{bmatrix} Y_1^T \\ Y_2^T \\ Y_3^T \\ \vdots \\ Y_T^T \end{bmatrix}, \quad \underset{T\times np}{X} = \begin{bmatrix} X_1^T \\ X_2^T \\ X_3^T \\ \vdots \\ X_T^T \end{bmatrix}, \quad \underset{T\times n}{A} = \begin{bmatrix} a_1^T \\ a_2^T \\ a_3^T \\ \vdots \\ a_T^T \end{bmatrix} \tag{6.33}$$

then we can rewrite the VAR(p) model in something resembling the linear regression form (i.e., $Y = X\beta + e$) as:

$$Y = X\phi + A. \tag{6.34}$$

This is not a simple linear regression since Y , ϕ, and A are not vectors but matrices. We can however convert it into a system of regression models.

Let $y^i = [Y_{it}]$ be the i^{th} column of Y which will be a $T \times 1$ vector of the i^{th} variable Y_{it} (e.g., money if $i = 2$ in (6.27)) and let the $T \times 1$ vector a^i be the i^{th} column of A, (e.g., the money shocks if $i = 2$ in (6.27)) so that:

$$\underset{T\times 1}{y^i} = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ \vdots \\ Y_{iT} \end{bmatrix}, \ \underset{T\times 1}{a^i} = \begin{bmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \\ \vdots \\ a_{iT} \end{bmatrix}.$$

Now let the $p \times 1$ vector β^i be the i^{th} column of ϕ. If for example $i = 2$ in (6.27) then β^2 would contain all coefficients in the money equation or:

$$\beta^2 = \begin{bmatrix} \phi_{21}^1 \\ \phi_{22}^1 \\ \phi_{21}^2 \\ \phi_{22}^2 \end{bmatrix}.$$

We have:

Theorem 205 *The VAR(p) model in* (6.34) *can be expressed as n regression models as:*

$$y^i = X\beta^i + a^i, \text{ for } i = 1, 2, \ldots n \tag{6.35}$$

where

$$a^i \sim N\left[0, \omega_{ii} I\right]$$

and

$$Cov\left[a^i, a^j\right] = E\left[a^i \left(a^j\right)^T\right] = \omega_{ij} I. \tag{6.36}$$

Remark 206 *Note that each regression has the same matrix of regressors X and that the error terms in the n regressions are contemporaneously correlated from* (6.36). *Thus* (6.35) *is a system of seemingly unrelated regressions called the SUR model.*

The ordinary least squares estimator of β^i is then:

$$\hat{\beta}^i = \left(X^T X\right)^{-1} X^T y^i \tag{6.37}$$

with asymptotic distribution

$$\sqrt{T}\left(\hat{\beta}^i - \beta^i\right) \overset{a}{\sim} N\left[0, \omega_{ii}\left(\operatorname*{plim}_{T\to\infty} \frac{X^T X}{T}\right)^{-1}\right]. \tag{6.38}$$

Thus it is legitimate to base hypothesis tests and confidence intervals on the estimated variance-covariance matrix:

$$\hat{Var}\left[\hat{\beta}^i\right] = \hat{\omega}_{ii}\left(X^T X\right)^{-1} \tag{6.39}$$

calculated by a regression packages. Furthermore if you wanted to test cross-equation restrictions then you could use:

$$\hat{Cov}\left[\hat{\beta}^i, \hat{\beta}^j\right] = \hat{\omega}_{ij}\left(X^T X\right)^{-1}$$

where $\hat{\omega}_{ij}$ is a consistent estimator of ω_{ij}.

6.2.2 Proof that OLS is Efficient

Ordinarily with a SUR model applying ordinary least squares to each equation individually results in inefficient estimates. There is however a striking result that when the regressors in each equation are identical then ordinary least squares and generalized least squares are numerically identical and so OLS is in fact efficient. Since for the VAR model the regressor matrix X in (6.37) is the same for all equations, this result applies. We will now proceed with the proof of this result, which is instructive because it illustrates the use of many of the mathematical tools that are important for multivariate analysis. We begin then with some preliminary mathematical results.

Some Mathematical Results

We often wish to write the elements of a matrix as a vector. If A is an $m \times n$ vector then $Vec\,[A]$ is a $mn \times 1$ column vector obtained by stacking the columns of A on top of each other. For example if

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

then

$$Vec\,[A] = \begin{bmatrix} 1 \\ 4 \\ 2 \\ 5 \\ 3 \\ 6 \end{bmatrix}. \tag{6.40}$$

The other important concept we will need is the Kronecker product. If A is $m \times n$ and B is $r \times s$ then the Kronecker product: $A \otimes B$ is an $mr \times ns$ matrix obtained by multiplying each element of $A = [a_{ij}]$ by B; that is

$$A \otimes B = [a_{ij} B] \tag{6.41}$$

For example if:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \tag{6.42}$$

then

$$A \otimes B = \begin{bmatrix} 1\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} & 2\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \\ \\ 3\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} & 4\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 5 & 6 & 10 & 12 \\ 7 & 8 & 14 & 16 \\ 15 & 18 & 20 & 24 \\ 21 & 24 & 28 & 32 \end{bmatrix} \tag{6.43}$$

We then have the following results:

Theorem 207

$$\begin{aligned} A \otimes (B + C) &= A \otimes B + A \otimes C \\ (A + B) \otimes C &= A \otimes C + B \otimes C \\ (A \otimes B)^T &= A^T \otimes B^T \\ (A \otimes B)(C \otimes D) &= AC \otimes BD \\ (A \otimes B)^{-1} &= A^{-1} \otimes B^{-1} \\ Vec\,[A + B] &= Vec\,[A] + Vec\,[B] \\ Vec\,[AB] &= (I \otimes A)\, Vec\,[B] \\ Vec\,[BC] &= \left(C^T \otimes I\right) Vec\,[B] \\ Vec\,[ABC] &= \left(C^T \otimes A\right) Vec\,[B]. \end{aligned}$$

Applying the $Vec\,[\,]$ operator to both sides of (6.34) and using the last result in Theorem 207 on

$$Y = X\phi + A \tag{6.44}$$

we obtain:

Theorem 208 *A VAR(p) model: $\phi(B)\,Y_t = a_t$ with T observations can be expressed as:*

$$y = (I \otimes X)\,\beta + a \tag{6.45}$$

where

$$y = Vec\,[Y]\,, \quad \beta = Vec\,[\phi]\,, \;\; a = Vec\,[A]\,. \tag{6.46}$$

We have:

Theorem 209 *The error term a has a variance-covariance matrix given by:*

$$E\left[aa^T\right] = \Omega \otimes I \tag{6.47}$$

where I here is a $T \times T$ identity matrix.

Proof. Since by definition:

$$a = \begin{bmatrix} a^1 \\ a^2 \\ \vdots \\ a^n \end{bmatrix}$$

and using (6.36) we obtain:

$$\begin{aligned} E\left[aa^T\right] &= E\begin{bmatrix} a^1\left(a^1\right)^T & a^1\left(a^2\right)^T & \cdots & a^1\left(a^n\right)^T \\ a^1\left(a^1\right)^T & a^1\left(a^1\right)^T & \cdots & a^1\left(a^1\right)^T \\ \vdots & \vdots & \ddots & \vdots \\ a^n\left(a^1\right)^T & a^n\left(a^2\right)^T & \cdots & a^n\left(a^n\right)^T \end{bmatrix} \\ &= \begin{bmatrix} \omega_{11}I & \omega_{12}I & \cdots & \omega_{1n}I \\ \omega_{12}I & \omega_{22}I & \cdots & \omega_{2n}I \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{1n}I & \omega_{2n}I & \cdots & \omega_{nn}I \end{bmatrix} \\ &= \Omega \otimes I. \end{aligned}$$

■

First consider estimating $\beta = \left[\beta^i\right]$ by the least squares estimator obtained by stacking $\hat{\beta}^i$ in a vector or:

$$\hat{\beta} = \left[\hat{\beta}^i\right] = \left[\left(X^T X\right)^{-1} X^T y^i\right].$$

This is equivalent to using (6.45) as a regression model with regression matrix $(I \otimes X)$ but ignoring the fact the variance covariance matrix of a is not of the form $\sigma^2 I$ but is $\Omega \otimes I$ by Theorem 209 . We therefore have:

$$\begin{aligned} \hat{\beta} &= \left((I \otimes X)^T (I \otimes X) \right)^{-1} (I \otimes X)^T y \\ &= \left(I \otimes \left(X^T X \right)^{-1} X^T \right) y. \end{aligned} \tag{6.48}$$

Now consider using the generalized least squares estimator $\tilde{\beta}$ where we use the fact that the variance covariance matrix of a is $\Omega \otimes I$ and so:

$$\tilde{\beta} = \left((I \otimes X)^T (\Omega \otimes I)^{-1} (I \otimes X) \right)^{-1} (I \otimes X) (\Omega \otimes I)^{-1} y. \tag{6.49}$$

We are now in a position to show that:

Theorem 210 *For the VAR(p) model ordinary and generalized least squares are identical or:*

$$\tilde{\beta} = \hat{\beta}$$

Proof. Using Theorem 207 we have:

$$\begin{aligned} \tilde{\beta} &= \left(\left(I \otimes X^T \right) \left(\Omega^{-1} \otimes I \right) (I \otimes X) \right)^{-1} (I \otimes X)^T (\Omega \otimes I)^{-1} y \\ &= \left(\Omega^{-1} \otimes \left(X^T X \right) \right)^{-1} \left(\Omega^{-1} \otimes X^T \right) y \\ &= \left(\Omega \otimes \left(X^T X \right)^{-1} \right) \left(\Omega^{-1} \otimes X^T \right) y \\ &= \left(I \otimes \left(X^T X \right)^{-1} X^T \right) y \\ &= \hat{\beta}. \end{aligned}$$

■

Note that since $\beta = Vec[\phi]$ we have:

$$\begin{aligned} Vec\left[\hat{\phi}\right] &= \left(I \otimes \left(X^T X \right)^{-1} X^T \right) y \\ &= \left(I \otimes \left(X^T X \right)^{-1} X^T \right) Vec[Y] \\ &= Vec\left[\left(X^T X \right)^{-1} X^T Y \right] \end{aligned} \tag{6.50}$$

from which it follows that:

Theorem 211 *The least squares estimator of ϕ is:*

$$\hat{\phi} = \left(X^T X \right)^{-1} X^T Y. \tag{6.51}$$

To estimate Ω we use:

Proposition 212 *A consistent estimator of* Ω *is*

$$\hat{\Omega} = \frac{\hat{A}^T \hat{A}}{T},$$

where:

$$\hat{A} = Y - X\hat{\phi}.$$

6.2.3 Maximum Likelihood

The log-likelihood for a VAR(p) model is given by:

$$l(\phi, \Omega) = -\frac{T}{2}\ln|\Omega| - \frac{1}{2}\sum_{t=1}^{T} a_t[\phi]^T \Omega^{-1} a_t[\phi] \tag{6.52}$$

where $|\Omega| \equiv \det[\Omega]$ and

$$a_t[\phi] = Y_t - \sum_{j=1}^{p} \phi_j Y_{t-j}. \tag{6.53}$$

Since GLS is the ML estimate under normality, and since we have shown that GLS and OLS are identical, it follows that

Theorem 213 *For a VAR(p) model the maximum likelihood estimate of* ϕ *is the least squares estimator:*

$$\hat{\phi}_{ML} = \hat{\phi} = \left(X^T X\right)^{-1} X^T Y \tag{6.54}$$

Now define the $n \times 1$ vector of least squares residuals $\hat{a}_t$ as:

$$\hat{a}_t = a_t\left[\hat{\phi}\right] = Y_t - \sum_{j=1}^{p} \hat{\phi}_j Y_{t-j}. \tag{6.55}$$

It then can be shown that

Proposition 214 *The* ML *estimate of* Ω *is given by*

$$\hat{\Omega} = \frac{1}{T}\sum_{t=1}^{T} \hat{a}_t \hat{a}_t^T = \frac{\hat{A}^T \hat{A}}{T} \tag{6.56}$$

If $\hat{\omega}_{ij}$ is the i, j^{th} of $\hat{\Omega}$ then this is equivalent to:

$$\hat{\omega}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \hat{a}_{it} \hat{a}_{jt} \tag{6.57}$$

where $\hat{a}_{it}$ is the i^{th} element of $\hat{a}_t$. Thus $\hat{\omega}_{ij}$ is the sample covariance between the least squares residuals in the i^{th} and j^{th} equations.

As before we will require the maximized log-likelihood:

$$l^* \equiv l\left(\hat{\phi}, \hat{\Omega}\right). \tag{6.58}$$

We have:

Theorem 215 *For a VAR(p) model the maximized log-likelihood is given by:*

$$l^* = -\frac{T}{2}\ln\left|\hat{\Omega}\right| - \frac{Tn}{2}.$$

Proof. Using the fact that the trace of a scalar is identical to that scalar it follows that:

$$\begin{aligned} l^* &= -\frac{T}{2}\ln\left|\hat{\Omega}\right| - \frac{1}{2}\sum_{t=1}^{T}\hat{a}_t^T\hat{\Omega}^{-1}\hat{a}_t \qquad (6.59) \\ &= -\frac{T}{2}\ln\left|\hat{\Omega}\right| - \frac{1}{2}\sum_{t=1}^{T}Tr\left[\hat{a}_t^T\hat{\Omega}^{-1}\hat{a}_t\right]. \end{aligned}$$

Using the fact that $Tr\left[AB\right] = Tr\left[BA\right]$ we have

$$\begin{aligned} l^* &= -\frac{T}{2}\ln\left|\hat{\Omega}\right| - \frac{1}{2}\sum_{t=1}^{T}Tr\left[\hat{a}_t^T\hat{\Omega}^{-1}\hat{a}_t\right] \\ &= -\frac{T}{2}\ln\left|\hat{\Omega}\right| - \frac{1}{2}\sum_{t=1}^{T}Tr\left[\hat{\Omega}^{-1}\hat{a}_t\hat{a}_t^T\right] \\ &= -\frac{T}{2}\ln\left|\hat{\Omega}\right| - \frac{1}{2}Tr\left[\hat{\Omega}^{-1}\underbrace{\sum_{t=1}^{T}\left(\hat{a}_t\hat{a}_t^T\right)}_{=T\hat{\Omega}}\right] \\ &= -\frac{T}{2}\ln\left|\hat{\Omega}\right| - \frac{T}{2}Tr\left[\hat{\Omega}^{-1}\hat{\Omega}\right] \\ &= -\frac{T}{2}\ln\left|\hat{\Omega}\right| - \frac{T}{2}Tr\left[\underset{n\times n}{I}\right] \\ &= -\frac{T}{2}\ln\left|\hat{\Omega}\right| - \frac{Tn}{2} \end{aligned}$$

■

6.2.4 Hypothesis Testing

We can use the formula for l^* to construct likelihood ratio test of restrictions on a $VAR(p)$ model. Given

$$\begin{aligned} H_0 &: \text{A Set of Restrictions versus} \\ H_1 &: \text{The Restrictions do not Hold} \end{aligned}$$

suppose we obtain the restricted log-likelihood:

$$l_R^* = -\frac{T}{2}\ln\left|\hat{\Omega}_R\right| - \frac{Tn}{2} \tag{6.60}$$

where $\hat{\Omega}_R$ is the estimator of Ω for the restricted model, and the unrestricted log-likelihood:

$$l_U^* = -\frac{T}{2}\ln\left|\hat{\Omega}_U\right| - \frac{Tn}{2} \tag{6.61}$$

where $\hat{\Omega}_U$ is the estimator of Ω for the unrestricted model.

Under H_0 we have the likelihood ratio statistic:

$$\Lambda = -2\left(l_R^* - l_U^*\right) = T\left(\ln\left|\hat{\Omega}_R\right| - \ln\left|\hat{\Omega}_U\right|\right) \tag{6.62}$$

so that:

$$\Lambda \overset{a}{\sim} \chi_r^2 \tag{6.63}$$

where r is the number of restrictions under H_o.

To be more concrete suppose we are testing

$$\begin{aligned} H_0 &: Y_t \sim VAR(p) \quad versus \\ H_1 &: Y_t \sim VAR(p+s). \end{aligned}$$

Then it follows that the likelihood ratio test statistic is:

$$\Lambda = T\left(\ln\left|\hat{\Omega}_p\right| - \ln\left|\hat{\Omega}_{p+s}\right|\right) \tag{6.64}$$

where $\hat{\Omega}_p$ is the estimate of Ω for the $VAR(p)$ model and $\hat{\Omega}_{p+s}$ is the estimate of Ω for the $VAR(p+s)$ model. Under H_0 we have

$$\Lambda \overset{a}{\sim} \chi_r^2 \tag{6.65}$$

where:

$$r = n^2 s. \tag{6.66}$$

Note that the number of restrictions is $n^2 s$ and not s as you might first think. This is because H_0 is

$$H_0 : \phi_{p+1} = \phi_{p+2} = \cdots = \phi_{p+s} = 0.$$

Recall that ϕ_j is an $n \times n$ matrix so that each ϕ_{p+j} has n^2 parameters.

6.2.5 Estimating p

We can also use the formula for l^* to derive the Akaike and Schwarz criteria for a $VAR(k)$ and use this in turn to estimate p.

The Akaike Information Criterion $AIC(K)$ for a VAR(k) is given by:

$$AIC(k) = \ln \left|\hat{\Omega}_k\right| + \frac{2n^2k}{T} \qquad k = 0, 1, 2, \ldots p_{\max}. \tag{6.67}$$

We can then estimate p by $\hat{p}$ the value of k which minimizes $AIC(k)$. As before $\hat{p}$ tends to choose overly large values of p. We have:

Proposition 216 *If $\hat{p}$ minimizes $AIC\,(k)$ then as $T \to \infty$*

$$plim\ \hat{p} \geq p \tag{6.68}$$

As before the Schwarz criterion $SC(k)$ given by:

$$SC(k) = \ln \left|\hat{\Omega}_k\right| + \frac{\ln(T)n^2k}{T}$$

remedies this defect of the AIC in that it picks the correct value of p asymptotically or

Proposition 217 *If $\hat{p}$ minimizes $SC(k)$ then as $T \to \infty$*

$$plim\ \hat{p} = p \tag{6.69}$$

6.3 Granger Causality

6.3.1 Defining Causality

Philosophers have long been arguing about the nature of causality and causality plays an important role in economic thought.

When estimating regressions it is easy to detect relations which reflect correlation and not causation. For example if we were to regress sun tan lotion sales on ice cream sale we would likely obtain a significant coefficient on the ice scream regressor. This would presumable reflect a correlation between the sales of ice cream and suntan lotion, but we would not want to conclude from this that ice cream consumption causes people to buy suntan lotion. Rather there is a third causal factor at work, weather, is causing ice cream and suntan lotion use to rise and fall together. It is particularly difficult to define causality with probabilistic models.

Granger causality is an attempt to make the notion of causality amenable to econometric analysis. Suppose we have two (possibly vector) time series Y_{1t} and Y_{2t} and we wish to determine which variable causes which.

One of the basic ideas of causality is that it is the past which causes the present and not vice versa. Furthermore, if say Y_{1t} causes Y_{2t} then we would expect past values of Y_{1t} to be useful in predicting present Y_{1t}.

Define I_{1t} as the information set which contains the past history of Y_{1t} , I_{2t} as the information set which contains the past history of Y_{2t} *so* that:

$$\begin{aligned} I_{1t} &= \{Y_{1t-1}, Y_{1t-1}, Y_{1t-3}, \dots\} \\ I_{2t} &= \{Y_{2t-1}, Y_{2t-1}, Y_{2t-3}, \dots\}. \end{aligned}$$

In addition define I_{3t} as the information set containing all other relevant information.

If Y_{1t} causes Y_{2t} then we would expect I_{1t} to be useful in predicting Y_{2t}. This leads to the following definitions:

Definition 218 ***Granger Causality*** *Y_{1t} Granger causes Y_{2t} or $Y_{1t} \rightarrow Y_{2t}$ if and only if:*

$$Var\left[Y_{2t}|I_{1t}, I_{2t}, I_{3t}\right] < Var\left[Y_{2t}|I_{2t}, I_{3t}\right].$$

Definition 219 *Y_{1t} does not Granger cause Y_{2t} or $Y_{1t} \nrightarrow Y_{2t}$ if and only if:*

$$Var\left[Y_{2t}|I_{1t}, I_{2t}, I_{3t}\right] = Var\left[Y_{2t}|I_{2t}, I_{3t}\right].$$

Thus Granger causality hinges on whether the past history of Y_{1t}, given by I_{1t}, is useful in predicting Y_{2t}.

Remark 220 *The presence of I_{3t} in the definition of Granger causality is there to insure that a variable is not helping in predicting simply because it is correlated with another variable that is causing the variable. For example suppose warm weather Y_{3t} causes both ice cream sales Y_{1t} and the number of stork nests Y_{2t} to rise. If past warm weather or I_{3t} were not included in the definition of Granger causality then one might mistakenly think that stork nests Y_{2t} are causing ice cream sales Y_{1t}.*

Remark 221 *In practice it is very difficult, if not impossible, to be sure that all of I_{3t} is included in any test of Granger causality.*

We can also talk about the causality going in the opposite direction; that is from Y_{2t} to Y_{1t} so that:

Definition 222 *Y_{2t} Granger causes Y_{1t} or $Y_{2t} \rightarrow Y_{1t}$ if and only if:*

$$Var\left[Y_{1t}|I_{1t}, I_{2t}, I_{3t}\right] < Var\left[Y_{1t}|I_{1t}, I_{3t}\right].$$

Definition 223 *Y_{2t} does not Granger cause Y_{1t} or $Y_{2t} \nrightarrow Y_{1t}$ if and only if:*

$$Var\left[Y_{1t}|I_{1t}, I_{2t}, I_{3t}\right] = Var\left[Y_{1t}|I_{1t}, I_{3t}\right].$$

Between any two time series then we can have four different situations:

1. No causality at all where:$Y_{1t} \nrightarrow Y_{2t}$ and $Y_{2t} \nrightarrow Y_{1t}$.
2. Unidirectional from Y_{1t} to Y_{2t} where: $Y_{1t} \rightarrow Y_{2t}$ and $Y_{2t} \nrightarrow Y_{1t}$,
3. Unidirectional from Y_{2t} to Y_{1t} where: $Y_{2t} \rightarrow Y_{1t}$ and $Y_{1t} \nrightarrow Y_{2t}$, and
4. Causality in both directions or *feedback* where: $Y_{1t} \rightarrow Y_{2t}$ and $Y_{2t} \rightarrow Y_{1t}$. s

6.3.2 Causality and Bivariate VAR(p)'s

Consider a bivariate $VAR(p)$ (i.e., with $n = 2$) so that

$$\begin{aligned} Y_{1t} &= \sum_{k=1}^{p} \phi_{11}^{k} Y_{1t-k} + \sum_{k=1}^{p} \phi_{12}^{k} Y_{2t-k} + a_{1t} \\ Y_{2t} &= \sum_{k=1}^{p} \phi_{21}^{k} Y_{1t-k} + \sum_{k=1}^{p} \phi_{22}^{k} Y_{2t-k} + a_{2t}. \end{aligned} \tag{6.70}$$

We can make Granger causality a statistically operational concept as follows. We have:

Theorem 224 *Given* (6.70) *then* $Y_{1t} \nrightarrow Y_{2t}$ *if and only if:*

$$\phi_{21}^{1} = \phi_{21}^{2} = \phi_{21}^{3} = \cdots = \phi_{21}^{p} = 0. \tag{6.71}$$

Theorem 225 *Given* (6.70) *then* $Y_{1t} \rightarrow Y_{2t}$ *if and only if:*

$$\phi_{21}^{1} \neq 0 \ \text{or} \ \phi_{21}^{2} \neq 0 \ \text{or} \ \cdots \ \text{or} \ \phi_{21}^{p} \neq 0. \tag{6.72}$$

Thus to test for causality we can make $Y_{1t} \nrightarrow Y_{2t}$ or that Y_{1t} does not Granger cause Y_{2t} the null hypothesis:

$$H_o : \phi_{21}^{1} = \phi_{21}^{2} = \phi_{21}^{3} = \cdots = \phi_{21}^{p} = 0 \ \left(Y_{1t} \nrightarrow Y_{2t}\right)$$

and test this against

$$H_1 : \phi_{21}^{1} \neq 0 \text{ or } \phi_{21}^{2} \neq 0 \text{ or } \cdots \text{ or } \phi_{21}^{p} \neq 0 \ \left(Y_{1t} \rightarrow Y_{2t}\right).$$

To perform this hypothesis test we run the unrestricted regression:

$$Y_{2t} = \sum_{k=1}^{p} \phi_{21}^{k} Y_{1t-k} + \sum_{k=1}^{p} \phi_{22}^{k} Y_{2t-k} + a_{2t}$$

and obtain the unrestricted estimate of the variance of $a_{2t} : \hat{\sigma}_{2U}^{2}$. We then run the restricted regression:

$$Y_{2t} = \sum_{k=1}^{p} \phi_{22}^{k} Y_{2t-k} + a_{2t}$$

and obtain the restricted estimate of the variance of $a_{2t} : \hat{\sigma}_{2R}^{2}$. Then under H_o :

$$\Lambda = T \ln\left(\frac{\hat{\sigma}_{2R}^{2}}{\hat{\sigma}_{2U}^{2}}\right) \overset{a}{\sim} \chi_{p}^{2}.$$

To test for causality in the opposite direction we have:

Theorem 226 *Given* (6.70) *then* $Y_{2t} \nrightarrow Y_{1t}$ *if and only if:*

$$\phi_{12}^{1} = \phi_{12}^{2} = \phi_{12}^{3} = \cdots = \phi_{12}^{p} = 0. \tag{6.73}$$

Theorem 227 *Given* (6.70) *then* $Y_{2t} \rightarrow Y_{1t}$ *if and only if:*

$$\phi_{12}^{1} \neq 0 \text{ or } \phi_{12}^{2} \neq 0 \text{ or } \cdots \text{ or } \phi_{12}^{p} \neq 0. \tag{6.74}$$

Thus to test for causality we can make $Y_{1t} \nrightarrow Y_{2t}$ or that Y_{1t} does not Granger cause Y_{2t} the null hypothesis so that:

$$H_o : \phi_{12}^{1} = \phi_{12}^{2} = \phi_{12}^{3} = \cdots = \phi_{12}^{p} = 0 \ \left(Y_{2t} \nrightarrow Y_{1t}\right)$$

and test this against

$$H_1 : \phi_{12}^{1} \neq 0 \text{ or } \phi_{12}^{2} \neq 0 \text{ or } \cdots \text{ or } \phi_{12}^{p} \neq 0. \ \left(Y_{2t} \rightarrow Y_{1t}\right).$$

To perform this hypothesis test we run the unrestricted regression:

$$Y_{1t} = \sum_{k=1}^{p} \phi_{11}^{k} Y_{1t-k} + \sum_{k=1}^{p} \phi_{12}^{k} Y_{2t-k} + a_{1t}$$

and obtain the unrestricted estimate of the variance of $a_{1t} : \hat{\sigma}_{1U}^2$. We then run the restricted regression:

$$Y_{1t} = \sum_{k=1}^{p} \phi_{11}^{k} Y_{1t-k} + a_{1t}$$

and obtain the restricted estimate of the variance of $a_{1t} : \hat{\sigma}_{1R}^2$. Then under H_o :

$$\Lambda = T \ln \left(\frac{\hat{\sigma}_{1R}^2}{\hat{\sigma}_{1U}^2} \right) \overset{a}{\sim} \chi_p^2.$$

6.3.3 An Economic Example

To make the preceding section more concrete, think of Y_{1t} as real GNP (appropriately detrended say as either a DS or TS model) and Y_{2t} as the money supply (again appropriately detrended).

As economists we might be interested in the relationship between Y_{1t} and Y_{2t}, that is the relationship between monetary policy and the business cycle. As such we would like to know if it is money that causes output or output that causes money.

Classical economic theory predicts that money is neutral; that money only affects nominal variables. Classical theory would therefore predict that $Y_{2t} \nrightarrow Y_{1t}$. If you look at the data, however, there is often found to be a correlation between money and output so to explain this correlation a classical theorist might argue that it is because money is endogenous; that increased economic activity causes increases in the money supply so that: $Y_{1t} \rightarrow Y_{2t}$.

Standard Keynesian or IS/LM analysis on the other hand predicts that changes in the money supply have an impact on GNP; for example when the money supply is increased the LM curves shifts to the right and GNP goes up. Thus a Keynesian economist would believe that $Y_{2t} \rightarrow Y_{1t}$.

Money is treated as an exogenous variable in the IS/LM model so that there is no causality from GNP to money. Thus a Keynesian economist would tend to believe that $Y_{1t} \nrightarrow Y_{2t}$.

We might therefore attempt to test:

$$\begin{aligned} H_0 &: \text{ money is neutral or } Y_{2t} \nrightarrow Y_{1t} \text{ versus} \\ H_1 &: \text{ money is not neutral or } Y_{2t} \rightarrow Y_{1t} \end{aligned}$$

A Keynesian would anticipate the rejection of H_o while the classical economist would anticipate accepting H_o.

Similarly we might test:

$$\begin{aligned} H_0 &: \text{ Money is Exogenous or } Y_{1t} \nrightarrow Y_{2t} \text{ versus} \\ H_1 &: \text{ Money is Endogenous or } Y_{1t} \rightarrow Y_{2t}. \end{aligned}$$

A Keynesian would now anticipate accepting H_o while the classical economist would anticipate rejecting H_o.

6.3.4 Instantaneous Causality

Definition

Generally speaking one thinks of causal mechanisms occurring across time with the past values causing the present. It is theoretically possible (but philosophically dubious) to have causality occurring instantaneously if adding say current Y_{1t} is useful in predicting current Y_{2t} . In particular we have:

Definition 228 ***Instantaneous Granger Causality*** *Y_{1t} Granger instantaneously causes Y_{2t} or $Y_{1t} \Leftrightarrow Y_{2t}$ if and only if:*

$$Var\left[Y_{2t}|Y_{1t}, I_{1t}, I_{2t}, I_{3t}\right] < Var\left[Y_{2t}|I_{1t}, I_{2t}, I_{3t}\right].$$

Definition 229 *Y_{1t} does not instantaneously Granger cause Y_{2t} or $Y_{1t} \not\Leftrightarrow Y_{2t}$ if and only if:*

$$Var\left[Y_{1t}|Y_{2t}, I_{1t}, I_{2t}, I_{3t}\right] = Var\left[Y_{1t}|I_{1t}, I_{2t}, I_{3t}\right].$$

Unlike ordinary causality, instantaneous causality has no direction. In particular:

Theorem 230 *$Y_{1t} \not\Leftrightarrow Y_{2t}$ if and only if $Y_{2t} \not\Leftrightarrow Y_{1t}$.*

For a VAR(p) model the existence of instantaneous causality depends on the variance-covariance matrix of a_t given by: Ω. In particular:

Theorem 231 *If Y_t follows a VAR(p) process then $Y_{it} \Longleftrightarrow Y_{jt}$ if and only if $\omega_{ij} \neq 0$. Alternatively $Y_{it} \not\Leftrightarrow Y_{jt}$ if and only if $\omega_{ij} = 0$.*

Thus a necessary and sufficient condition for instantaneous causality between Y_{it} and Y_{jt} to exist is that a_{it} and a_{jt} be contemporaneously correlated with each other.

Proof. Recall that for any two scalar random variables: X_1 and X_2 that are jointly normally distributed that:

$$\begin{aligned} E[X_1|X_2] &= E[X_1] + \frac{Cov[X_1, X_2]}{Var[X_2]}(X_2 - E[X_2]) \\ Var[X_1|X_2] &= Var[X_1]\left(1 - \frac{Cov[X_1, X_2]^2}{Var[X_1]\,Var[X_2]}\right). \end{aligned}$$

Now let X_1 be Y_{it} and let X_2 be Y_{jt}. Then:

$$\begin{aligned} Var_{t-1}[Y_{it}|Y_{jt}] &= Var_{t-1}[Y_{it}]\left(1 - \frac{Cov_{t-1}[Y_{it}, Y_{jt}]^2}{Var_{t-1}[Y_{it}]\,Var_{t-1}[Y_{jt}]}\right) \\ &= Var_{t-1}[Y_{it}]\left(1 - \frac{\omega_{ij}^2}{\omega_{ii}\omega_{jj}}\right) \end{aligned}$$

since for a VAR(p)

$$Cov_{t-1}[Y_{it}, Y_{jt}] = \omega_{ij}.$$

Thus $Var_{t-1}[Y_{it}|Y_{jt}] = Var_{t-1}[Y_{it}]$ if and only if $\omega_{ij} = 0$. ■

The Structural VAR Representation

Normally we write a VAR(p) as:

$$Y_t = \sum_{k=1}^{p} \phi_k Y_{t-k} + a_t. \tag{6.75}$$

and so there are no contemporaneous values of Y_t on the right-hand side to act as regressors. There is an alternative representation of a VAR(p) where contemporaneous Y_t *'s* do appear on the right-hand side but where the error terms are contemporaneously uncorrelated. In particular we have:

Theorem 232 ***Structural VAR Representation****: The VAR(p) model in* (6.75) *has an equivalent representation:*

$$Y_t = \tilde{\phi}_o Y_t + \sum_{k=1}^{p} \tilde{\phi}_k Y_{t-k} + \varepsilon_t$$

where $\tilde{\phi}_o$ is a upper triangular matrix[3] *with zeros along the diagonal and*

$$\varepsilon_t \sim N[0, \Lambda]$$

where Λ is a diagonal matrix (or $Cov[\varepsilon_{it}, \varepsilon_{jt}] = 0$ for $i \neq j$).

Proof. Using the Cholesky decomposition we can write the variance-covariance matrix of a_t as:

$$\Omega = C \Lambda C^T$$

where C is an upper triangular matrix with $1's$ along the diagonal and Λ is a diagonal matrix with positive elements along the diagonal. Multiplying both sides of (6.75) by C^{-1} we obtain:

$$C^{-1} Y_t = \sum_{k=1}^{p} C^{-1} \phi_k Y_{t-k} + C^{-1} a_t.$$

Define:

$$\varepsilon_t = C^{-1} a_t \sim N[0, \Lambda]$$

(so that $Cov[\varepsilon_{it}, \varepsilon_{jt}] = 0$ for $i \neq j$) and

$$\begin{aligned} \tilde{\phi}_o &= I - C^{-1} \\ \tilde{\phi}_k &= C^{-1} \phi_k. \end{aligned}$$

Since C^{-1} is upper triangular with $1's$ along the diagonal, $\tilde{\phi}_o = I - C^{-1}$ is upper triangular with $0's$ along the diagonal. Now using $C^{-1} Y_t = Y_t - \tilde{\phi}_o Y_t$ and putting $\tilde{\phi}_o Y_t$ on the right-hand side we obtain:

$$Y_t = \tilde{\phi}_o Y_t + \sum_{k=1}^{p} \tilde{\phi}_k Y_{t-k} + \varepsilon_t.$$

■

An Example

To make this more concrete consider the case where $n = 2$ so that:

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = \sum_{k=1}^{p} \begin{bmatrix} \phi_{11}^k & \phi_{12}^k \\ \phi_{21}^k & \phi_{22}^k \end{bmatrix} \begin{bmatrix} Y_{1t-k} \\ Y_{2t-k} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}. \tag{6.76}$$

Given the variance-covariance matrix of a_t :

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{bmatrix}$$

[3] That is all elements below the diagonal are zero.

the Cholesky decomposition is:

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{bmatrix} = \begin{bmatrix} 1 & \beta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix}.$$

You can verify this by multiplying out the right-hand side to show that:

$$\begin{aligned} \beta &= \frac{\omega_{12}}{\omega_{22}},\ C = \begin{bmatrix} 1 & \frac{\omega_{12}}{\omega_{22}} \\ 0 & 1 \end{bmatrix}, \qquad (6.77) \\ \lambda_1 &= \omega_{11} - \frac{\omega_{12}^2}{\omega_{22}},\ \lambda_2 = \omega_{22}. \end{aligned}$$

Thus

$$C^{-1} = \begin{bmatrix} 1 & -\beta \\ 0 & 1 \end{bmatrix}$$

and so multiplying both sides of (6.76) by C^{-1} as:

$$\begin{bmatrix} 1 & -\beta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = \sum_{k=1}^{p} \begin{bmatrix} 1 & -\beta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \phi_{11}^k & \phi_{12}^k \\ \phi_{21}^k & \phi_{22}^k \end{bmatrix} \begin{bmatrix} Y_{1t-k} \\ Y_{2t-k} \end{bmatrix} + \begin{bmatrix} 1 & -\beta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}$$

and putting $-\beta Y_{2t}$ on the right-hand side we obtain

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} + \sum_{k=1}^{p} \begin{bmatrix} \tilde{\phi}_{11}^k & \tilde{\phi}_{12}^k \\ \tilde{\phi}_{21}^k & \tilde{\phi}_{22}^k \end{bmatrix} \begin{bmatrix} Y_{1t-k} \\ Y_{2t-k} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

or

$$\begin{aligned} Y_{1t} &= \beta Y_{2t} + \sum_{k=1}^{p} \tilde{\phi}_{11}^k Y_{1t-k} + \sum_{k=1}^{p} \tilde{\phi}_{12}^k Y_{2t-k} + \varepsilon_{1t} \qquad (6.78) \\ Y_{2t} &= \sum_{k=1}^{p} \tilde{\phi}_{21}^k Y_{1t-k} + \sum_{k=1}^{p} \tilde{\phi}_{22}^k Y_{2t-k} + \varepsilon_{2t} \end{aligned}$$

where

$$\tilde{\phi}_o = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix}$$

and

$$\tilde{\phi}_k = \begin{bmatrix} \tilde{\phi}_{11}^k & \tilde{\phi}_{12}^k \\ \tilde{\phi}_{21}^k & \tilde{\phi}_{22}^k \end{bmatrix} = \begin{bmatrix} \phi_{11}^k - \beta\phi_{21}^k & \phi_{12}^k - \beta\phi_{22}^k \\ \phi_{21}^k & \phi_{22}^k \end{bmatrix}$$

and

$$\varepsilon_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} = \begin{bmatrix} a_{1t} - \beta a_{2t} \\ a_{2t} \end{bmatrix}.$$

Note that $\tilde{\phi}_o$ is upper triangular with zeros along the diagonal. You might want to verify directly that $Cov\left[\varepsilon_{1t}, \varepsilon_{2t}\right] = 0$.

Testing

There are two ways of testing for instantaneous causality. I will only consider the case where $n = 2$ but generalizations are straightforward.

First we can test:

$$\begin{aligned} H_0 &: \omega_{12} = 0 \text{ (or } Y_{1t} \not\Leftrightarrow Y_{2t} \text{) versus} \\ H_1 &: \omega_{12} \neq 0 \text{ (or } Y_{1t} \Leftrightarrow Y_{2t} \text{)} \end{aligned}$$

using a likelihood ratio test. Under H_0 the restricted maximum likelihood estimator of Ω is:

$$\hat{\Omega}_R = \begin{bmatrix} \hat{\omega}_{11} & 0 \\ 0 & \hat{\omega}_{22} \end{bmatrix} \tag{6.79}$$

where $\hat{\omega}_{11}$ and $\hat{\omega}_{22}$ are the estimated variances from running the two regressions:

$$\begin{aligned} Y_{1t} &= \sum_{k=1}^{p} \phi_{11}^{k} Y_{1t-k} + \sum_{k=1}^{p} \phi_{12}^{k} Y_{2t-k} + a_{1t} \\ Y_{2t} &= \sum_{k=1}^{p} \phi_{21}^{k} Y_{1t-k} + \sum_{k=1}^{p} \phi_{22}^{k} Y_{2t-k} + a_{2t} \end{aligned} \tag{6.80}$$

or:

$$\hat{\omega}_{11} = \frac{1}{T}\sum_{t=1}^{T} \hat{a}_{1t}^2 \ , \quad \hat{\omega}_{22} = \frac{1}{T}\sum_{t=1}^{T} \hat{a}_{2t}^2. \tag{6.81}$$

The unrestricted estimator of Ω is

$$\hat{\Omega}_U = \begin{bmatrix} \hat{\omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{12} & \hat{\omega}_{22} \end{bmatrix} \tag{6.82}$$

where

$$\hat{\omega}_{12} = \frac{1}{T}\sum_{t=1}^{T} \hat{a}_{1t}\hat{a}_{2t} \tag{6.83}$$

is the estimated correlation between a_{1t} and a_{2t} from the above two regression and $\hat{\omega}_{11}$ and $\hat{\omega}_{22}$ are the same as before (this follows since we showed that OLS and GLS are the same). The likelihood ratio statistic is then:

$$\begin{aligned} \Lambda &= T\left(\ln\left|\hat{\Omega}_R\right| - \ln\left|\hat{\Omega}_U\right|\right) \\ &= -T\ln\left(1 - \hat{\rho}_{12}^2\right) \end{aligned} \tag{6.84}$$

where

$$\hat{\rho}_{12} = \frac{\hat{\omega}_{12}}{\sqrt{\hat{\omega}_{11}}\sqrt{\hat{\omega}_{22}}} \tag{6.85}$$

is the estimated correlation between a_{1t} and a_{2t}. Under H_0 then $\Lambda \overset{a}{\sim} \chi_1^2$.

The second method for testing instantaneous causality is to test

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta \neq 0$$

using the first equation of the structural VAR model:

$$Y_{1t} = \beta Y_{2t} + \sum_{k=1}^{p} \tilde{\phi}_{11}^{k} Y_{1t-k} + \sum_{k=1}^{p} \tilde{\phi}_{12}^{k} Y_{2t-k} + \varepsilon_{1t} \tag{6.86}$$

and doing a t test on $\hat{\beta}$ in the usual way.

6.4 Forecasting

6.4.1 Theory

As with ARMA(p,q) models we can recursively calculate forecasts for a multivariate time series Y_t and construct confidence intervals. Here I will only discuss the $VAR(p)$ model.

To calculate forecasts: $E_t[Y_{t+k}]$ recursively we use:

Theorem 233 *For an VAR(p) process:*

$$E_t[Y_{t+k}] = \phi_1 E_t[Y_{t+k-1}] + \phi_2 E_t[Y_{t+k-2}] + \cdots + \phi_p E_t[Y_{t+k-p}] \tag{6.87}$$

with starting values given by:

$$E_t[Y_{t+k}] = Y_{t+k} \quad for \quad k \leq 0. \tag{6.88}$$

The problem now is to calculate confidence intervals for our forecasts. We use the following result.

Theorem 234 *For a VAR(p):*

$$Var_t[Y_{t+k}] = \Omega + \psi_1 \Omega \psi_1^T + \psi_2 \Omega \psi_2^T + \cdots + \psi_{k-1} \Omega \psi_{k-1}^T$$

where ψ_k is calculated recursively from:

$$\psi_k = \phi_1 \psi_{k-1} + \phi_2 \psi_{k-2} + \cdots + \phi_p \psi_{k-p}$$

and with starting values $\psi_0 = I$ and $\psi_k = 0$ for $k < 0$.

We can calculate $Var_t[Y_{t+k}]$ recursively using:

Theorem 235 *For a VAR(p):*

$$Var_t[Y_{t+k}] = Var_t[Y_{t+k-1}] + \psi_{k-1} \Omega \psi_{k-1}^T \tag{6.89}$$

with:

$$Var_t[Y_{t+1}] = \Omega.$$

Suppose we wish to forecast some linear combination of the elements of Y_{t+k}, say $c^T Y_{t+k}$ where c is an $n \times 1$ vector. We have:

Theorem 236 *A 95% confidence interval for $c^T Y_{t+k}$ is:*

$$c^T E_t[Y_{t+k}] \pm 1.96\sqrt{c^T Var_t\left[Y_{t+k}\right]c}.$$

Note that if we wish to forecast say Y_{it+k} we simply set the i^{th} element of c to 1 and all other elements equal to zero in which case we have:

Theorem 237 *A 95% confidence interval for Y_{it+k} is:*

$$E_t[Y_{it+k}] \pm 1.96\sqrt{Var_t\left[Y_{it+k}\right]}$$

where $E_t[Y_{it+k}]$ is the i^{th} element of $E_t[Y_{t+k}]$ and $Var_t\left[Y_{it+k}\right]$ is the i^{th} diagonal element of $Var_t\left[Y_{t+k}\right]$.

Allowing arbitrary linear combinations when forecasting may be useful. For example suppose that $n = 2$ and $Y_{1t} = \ln(UN_t)$ where UN_t is the number of unemployed and $Y_{2t} = \ln(LF_t)$ where LF_t is the labour force. Then if:

$$c^T = \left[\begin{array}{cc} 1 & -1 \end{array}\right]$$

then

$$c^T Y_t = \left[\begin{array}{cc} 1 & -1 \end{array}\right]\left[\begin{array}{c} Y_{1t} \\ Y_{2t} \end{array}\right] = Y_{1t} - Y_{2t} = \ln\left(\frac{UN_t}{LF_t}\right)$$

and so $c^T Y_t$ is the log of the rate of unemployment. Thus we could use Theorem 236 to construct a forecast for the rate of unemployment.

6.4.2 A Worked Example

An ordinary $AR(p)$ model uses only a series own past history to construct forecasts. A $VAR(p)$ on the other hand uses in addition the past history of other series and so one might expect in many situations that this will provide better forecasts. For example it has been found that such variables as the term structure of interest rates are very useful when forecasting the business cycle.

As usual the matrix notation for forecasting a VAR(p) hides a lot of details that are important. In general you will want to use the computer to do the tedious work of calculating forecasts and constructing confidence intervals but to do the programming properly you need to understand what is going on inside the matrices.

Suppose then we have $n = 2$ time series and we have estimated a VAR(2)

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + a_t$$

as:

$$\begin{aligned} Y_{1t} &= 0.5Y_{1t-1} - 0.1Y_{2t-1} + 0.2Y_{1t-1} + 0.3Y_{2t-1} + a_{1t} \\ Y_{2t} &= 0.5Y_{1t-1} + 0.3Y_{2t-1} + 0.1Y_{1t-1} + 0.2Y_{2t-1} + a_{2t}. \end{aligned} \tag{6.90}$$

To make things concrete you may want to think of Y_{1t} as say detrended GNP and Y_{2t} as detrended money.

In matrix notation ϕ_1 and ϕ_2 are given by:

$$\phi_1 = \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}, \ \phi_2 = \begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}.$$

Suppose further that the estimated variance-covariance matrix is:

$$\Omega = \begin{bmatrix} (0.05)^2 & 0.0001 \\ 0.0001 & (0.08)^2 \end{bmatrix}$$

and we observe at time t that:

$$\begin{aligned} Y_t &= \begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = \begin{bmatrix} 0.03 \\ 0.05 \end{bmatrix} \\ Y_{t-1} &= \begin{bmatrix} Y_{1t-1} \\ Y_{2t-1} \end{bmatrix} = \begin{bmatrix} 0.04 \\ 0.06 \end{bmatrix}. \end{aligned} \tag{6.91}$$

To calculate the one-step ahead forecast: $E_t\left[Y_{t+1}\right]$ we use:

$$\begin{aligned} E_t\left[Y_{1t+1}\right] &= 0.5E_t\left[Y_{1t}\right] - 0.1E_t\left[Y_{2t}\right] + 0.2E_t\left[Y_{1t-1}\right] + 0.3E_t\left[Y_{2t-1}\right] \\ E_t\left[Y_{2t+1}\right] &= 0.5E_t\left[Y_{1t}\right] + 0.3E_t\left[Y_{1t}\right] + 0.1E_t\left[Y_{1t-1}\right] + 0.2E_t\left[Y_{2t-1}\right] \end{aligned}$$

or in matrix notation

$$E_t\left[Y_{t+1}\right] = \phi_1 E_t\left[Y_t\right] + \phi_2 E_t\left[Y_{t-2}\right]$$

or:

$$\begin{bmatrix} E_t\left[Y_{1t+1}\right] \\ E_t\left[Y_{2t+1}\right] \end{bmatrix} = \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix} \begin{bmatrix} E_t\left[Y_{1t}\right] \\ E_t\left[Y_{2t}\right] \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} E_t\left[Y_{1t-1}\right] \\ E_t\left[Y_{2t-1}\right] \end{bmatrix}.$$

From the starting values in (6.91) we then have:

$$\begin{aligned} \begin{bmatrix} E_t\left[Y_{1t+1}\right] \\ E_t\left[Y_{2t+1}\right] \end{bmatrix} &= \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix} \begin{bmatrix} 0.03 \\ 0.05 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} 0.04 \\ 0.06 \end{bmatrix} \\ &= \begin{bmatrix} .036 \\ .046 \end{bmatrix}. \end{aligned}$$

Thus our one-step ahead forecasts are $E_t\left[Y_{1t+1}\right] = 0.036$ or 3.6% for GNP growth and $E_t\left[Y_{2t+1}\right] = 0.046$ or 4.6% for money growth.

To calculate the two step ahead forecast we repeat the procedure using:

$$E_t\left[Y_{t+2}\right] = \phi_1 E_t\left[Y_{t+1}\right] + \phi_2 E_t\left[Y_t\right]$$

or:

$$\begin{aligned}\begin{bmatrix} E_t\left[Y_{1t+2}\right] \\ E_t\left[Y_{2t+2}\right] \end{bmatrix} &= \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}\begin{bmatrix} 0.036 \\ 0.046 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}\begin{bmatrix} 0.03 \\ 0.05 \end{bmatrix} \\ &= \begin{bmatrix} 0.0344 \\ 0.0448 \end{bmatrix}.\end{aligned}$$

Thus our two-step ahead forecasts are $E_t\left[Y_{1t+2}\right] = 0.0344$ and $E_t\left[Y_{2t+1}\right] = 0.0448$.

Finally let us calculate the three-step ahead forecast as:

$$\begin{aligned}\begin{bmatrix} E_t\left[Y_{1t+3}\right] \\ E_t\left[Y_{2t+3}\right] \end{bmatrix} &= \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}\begin{bmatrix} 0.0344 \\ 0.0448 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}\begin{bmatrix} 0.036 \\ 0.046 \end{bmatrix} \\ &= \begin{bmatrix} 0.03372 \\ 0.04344 \end{bmatrix}.\end{aligned}$$

To construct confidence intervals we first need to calculate the $\psi_k\ 's$. We have as starting values for the recursive calculations:

$$\psi_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \psi_{-1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

To calculate ψ_1 we then use:

$$\psi_1 = \phi_1\psi_0 + \phi_2\psi_{-1}$$

or:

$$\begin{aligned}\psi_1 &= \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}.\end{aligned}$$

Then to calculate ψ_2 we use:

$$\psi_2 = \phi_1\psi_1 + \phi_2\psi_0$$

or:

$$\begin{aligned}\psi_2 &= \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}\begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.4 & 0.22 \\ 0.5 & 0.24 \end{bmatrix}.\end{aligned}$$

Finally let us calculate ψ_3 and ψ_4 as

$$\begin{aligned}\psi_3 &= \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}\begin{bmatrix} 0.4 & 0.22 \\ 0.5 & 0.24 \end{bmatrix}+\begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}\begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix} \\ &= \begin{bmatrix} 0.4 & 0.156 \\ 0.5 & 0.232 \end{bmatrix}\end{aligned}$$

and:

$$\begin{aligned}\psi_4 &= \begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}\begin{bmatrix} 0.4 & 0.156 \\ 0.5 & 0.232 \end{bmatrix}+\begin{bmatrix} 0.2 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}\begin{bmatrix} 0.4 & 0.22 \\ 0.5 & 0.24 \end{bmatrix} \\ &= \begin{bmatrix} 0.38 & 0.1708 \\ 0.49 & 0.2176 \end{bmatrix}.\end{aligned}$$

This kind of calculation is something that computers are obviously very good at.

Now consider calculating the conditional variance-covariance matrix of Y_{t+k}. Using (6.89) we have first that:

$$Var_t\left[Y_{t+1}\right]=\Omega=\begin{bmatrix} (0.05)^2 & 0.0001 \\ 0.0001 & (0.08)^2 \end{bmatrix}.$$

To calculate $Var_t\left[Y_{t+2}\right]$ we then use:

$$Var_t\left[Y_{t+2}\right]=Var_t\left[Y_{t+1}\right]+\psi_1\Omega\psi_1^T$$

or

$$\begin{aligned}Var_t\left[Y_{t+2}\right] &= \begin{bmatrix} (0.05)^2 & 0.0001 \\ 0.0001 & (0.08)^2 \end{bmatrix} \\ &\quad+\begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}\begin{bmatrix} (0.05)^2 & 0.0001 \\ 0.0001 & (0.08)^2 \end{bmatrix}\begin{bmatrix} 0.5 & -0.1 \\ 0.5 & 0.3 \end{bmatrix}^T \\ &= \begin{bmatrix} 3.179\times10^{-3} & 5.43\times10^{-4} \\ 5.43\times10^{-4} & 7.631\times10^{-3} \end{bmatrix}.\end{aligned}$$

Then to calculate $Var_t\left[Y_{t+3}\right]$ we use:

$$\begin{aligned}Var_t\left[Y_{t+3}\right] &= Var_t\left[Y_{t+2}\right]+\psi_2\Omega\psi_2^T \\ &= \begin{bmatrix} 3.179\times10^{-3} & 5.43\times10^{-4} \\ 5.43\times10^{-4} & 7.631\times10^{-3} \end{bmatrix} \\ &\quad+\begin{bmatrix} 0.4 & 0.22 \\ 0.5 & 0.24 \end{bmatrix}\begin{bmatrix} (0.05)^2 & 0.0001 \\ 0.0001 & (0.08)^2 \end{bmatrix}\begin{bmatrix} 0.4 & 0.22 \\ 0.5 & 0.24 \end{bmatrix}^T \\ &= \begin{bmatrix} 3.9064\times10^{-3} & 1.4015\times10^{-3} \\ 1.4015\times10^{-3} & 8.6486\times10^{-3} \end{bmatrix}.\end{aligned}$$

A 95% confidence interval for Y_{1t+3} is then given by:

$$E_t\left[Y_{1t+3}\right]\pm1.96\sqrt{Var_t\left[Y_{1t+3}\right]}$$

where $Var_t[Y_{1t+3}]$ is the $(1,1)$ element of $Var_t[Y_{t+3}]$ above. The confidence interval is therefore:

$$0.03372 \pm 1.96\sqrt{3.9064 \times 10^{-3}}$$

or

$$0.03372 \pm 0.1225.$$

Similarly a 95% confidence interval for Y_{2t+3} is given by:

$$0.04344 \pm 1.96\sqrt{8.6486 \times 10^{-3}}$$

or

$$0.03372 \pm 0.18228.$$

If we wished to calculate a confidence interval for a linear combination, say:

$$W_{t+3} = 0.6Y_{1t+3} + 0.4Y_{2t+3}$$

or $W_{t+3} = c^T Y_{t+3}$ where:

$$c^T = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}$$

we would use:

$$c^T E_t[Y_{t+3}] \pm 1.96\sqrt{c^T Var_t[Y_{t+3}]\, c}$$

or:

$$\begin{bmatrix} 0.6 & 0.4 \end{bmatrix} \begin{bmatrix} 0.03372 \\ 0.04344 \end{bmatrix} \pm 1.96\sqrt{\begin{bmatrix} 0.6 & 0.4 \end{bmatrix} \begin{bmatrix} 3.9064 \times 10^{-3} & 1.4015 \times 10^{-3} \\ 1.4015 \times 10^{-3} & 8.6486 \times 10^{-3} \end{bmatrix} \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix}}$$

which reduces to:

$$0.0376 \pm 0.115.$$

6.5 Cointegration

6.5.1 The Long-Run in Economics

Many times in economics we know that a theory is false in the short-run but still believe it to be true in the long-run. Consider for example the theory of purchasing power parity (PPP) which states that:

$$P_t = e_t P_t^F \tag{6.92}$$

where P_t is the price of some domestic good, P_t^F is the price of the same good in a foreign country, and e_t is the exchange rate between the two countries.

We know that PPP does not hold in the short-run. Generally speaking in 1999 if you go from Canada to the United States you will find goods more expensive than in Canada while in 1989 you would find the same goods cheaper in the United States. Nevertheless many economists still believe that PPP holds in the long-run.

Many economic models make balanced growth predictions; for example that consumption C_t or investment I_t should be a constant proportion of national income: Y_t or:

$$\begin{aligned} C_t &= \beta Y_t \qquad (6.93)\\ I_t &= \gamma Y_t. \end{aligned}$$

Clearly these predictions are false since we observe that the ratios of consumption and investment to GNP vary over time. Nevertheless these ratios do not change by very much and so one is tempted to say that these balanced growth relationships hold in the long run even though they do not hold in the short run.

Finally consider a demand for money or LM relationship of the form:

$$M_t^s = AY_t^\alpha R_t^\beta P_t^\gamma \qquad (6.94)$$

where M_t^s is the money supply, P_t is the price level, Y_t is real GNP and R_t is an interest rate. Again we know that such a relationship does not hold exactly in the short-run, say because of short-run money market dynamics, but we may nevertheless still believe that such a relationship will hold over the long-run.

Cointegration provides a theoretical framework with which we can give a precise meaning to the idea of a relationship holding in the long-run, and which allow us to estimate and test these long-run relationships.

6.5.2 Some Theory

To begin we need some definitions and results commonly used in the cointegration literature.

Definition 238 *We say that Z_t is integrated of order d or*

$$Z_t \sim I(d)$$

if $(1-B)^{d-1} Z_t$ is nonstationary and $(1-B)^d Z_t$ is stationary and invertible. If Z_t is stationary we say that Z_t is $I(0)$.

For our purposes we only need to refer to $I(0)$ or stationary series and $I(1)$ or series that have a unit root. $I(0)$ and $I(1)$ series behave very differently. An $I(0)$ series is stationary and hence fluctuates around its mean value. If that mean value is zero then an $I(0)$ series will generally speaking not wander too far from zero.

An $I(1)$ series on the other hand is nonstationary. It behaves like a random walk (and a random walk is $I(1)$) in that it does not have a mean value around which it fluctuates. $I(1)$ series wander a great deal. Roughly speaking after k periods have elapsed an $I(1)$ series will have wandered distance proportional to $\sqrt{k}$ and so will eventually move an unbounded distance from zero.

We have the following results for $I(0)$ and $I(1)$ variables:

Theorem 239 *If $Z_t \sim I(0)$ with $E[Z_t] = 0$ then Z_t will cross the horizontal axis (go from being positive to negative or from negative to positive) an infinite number of times with probability* 1.

Theorem 240 *If $Z_t \sim I(1)$ and $E[\Delta Z_t] = 0$ then Z_t will cross the horizontal axis with probability* 1 *but only a finite number of times. Furthermore the expected length of time between crossings is infinite.*

We have the following results for $I(0)$ and $I(1)$ variables:

Theorem 241 *If $X_t \sim I(0)$ and $Y_t \sim I(0)$ then for any nonrandom constants α and β:*

$$\alpha X_t + \beta Y_t \sim I(0)$$

Theorem 242 *If $X_t \sim I(1)$ and $Y_t \sim I(0)$ then:*

$$X_t + Y_t \sim I(1).$$

Thus adding two stationary series together always yields another stationary series while adding a stationary series to a non-stationary series always yields a nonstationary series.

Normally one would expect that adding two $I(1)$ series together can would lead to another $I(1)$; that is if $X_t \sim I(1)$ and $Y_t \sim I(1)$ then

$$\alpha X_t + \beta Y_t \sim I(1).$$

However, it turns out to be possible that adding two $I(1)$ series together can lead to an $I(0)$ series; that is:

$$\alpha X_t + \beta Y_t \sim I(0).$$

This occurs when X_t and Y_t are cointegrated and α and β then form the components of a cointegrating vector.

More formally the definition of cointegration is as follows:

Definition 243 *If for the $n \times 1$ vector time series $X_t \sim I(1)$ there exists a nonrandom $n \times 1$ vector $c \neq 0$ such that:*

$$Z_t \equiv c^T X_t \sim I(0)$$

then we say that X_t is cointegrated with cointegrating vector c.

Remark 244 *If $c = 0$ then $c^T X_t = 0$, which is in a trivial sense always stationary or $I(0)$. This is the reason we impose the requirement that $c \neq 0$.*

Actually, if X_t has one cointegrating vector then in a trivial sense it has an infinite number of cointegrating vectors. Thus it will be necessary to make our questions more precise if we are to make them meaningful. This is because of the following two theorems:

Theorem 245 *If X_t has a cointegrating vector c then if α is any nonzero constant then*

$$d = \alpha c$$

is also a cointegrating vector.

Proof. Given that:

$$c^T X_t \sim I(0)$$

if we multiply both sides by α we then find that:

$$d^T X_t = \alpha c^T X_t \sim I(0) \tag{6.95}$$

since a scalar multiplied by a stationary $I(0)$ series is also $I(0)$. We therefore conclude that d is also a cointegrating vector. ■

The problem of non-uniqueness here could be easily fixed by requiring say that the first non-zero element in c be 1 and so we could rule out d as a cointegrating vector.

However we also have:

Theorem 246 *Suppose X_t has two cointegrating vectors c_1 and c_2. Then $c = c_1 + c_2$ is also a cointegrating vector.*

Proof. This follows since:

$$c^T X_t = (c_1 + c_2)^T X_t = c_1^T X_t + c_2^T X_t \sim I(0) \tag{6.96}$$

from Theorem 241. ■

Combining these two results we see that in general if c_1 and c_2 are two cointegrating vectors then so is

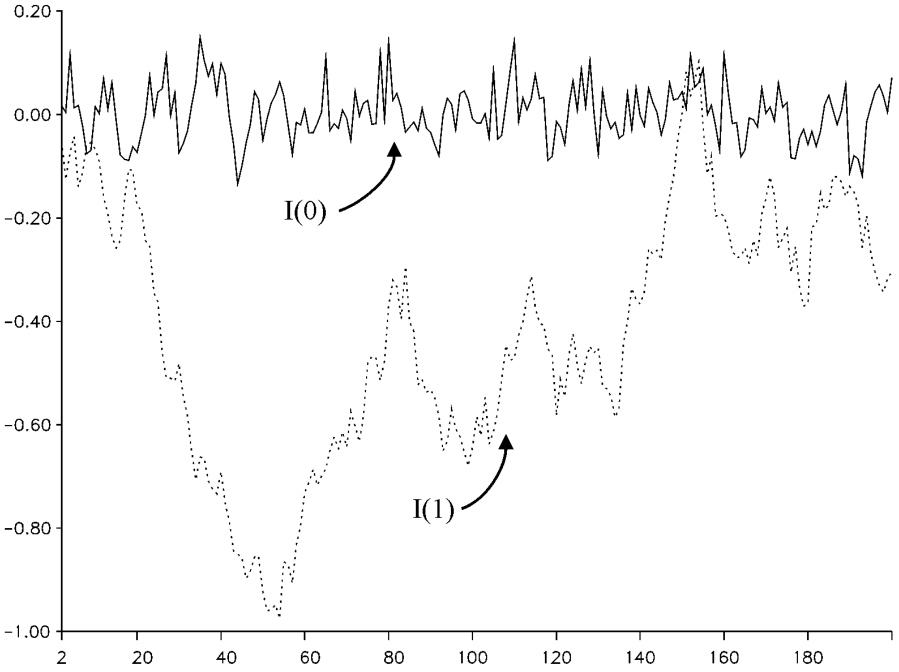
$$\alpha_1 c_1 + \alpha_2 c_2 \tag{6.97}$$

where α_1 and α_2 are two scalars and so we have:

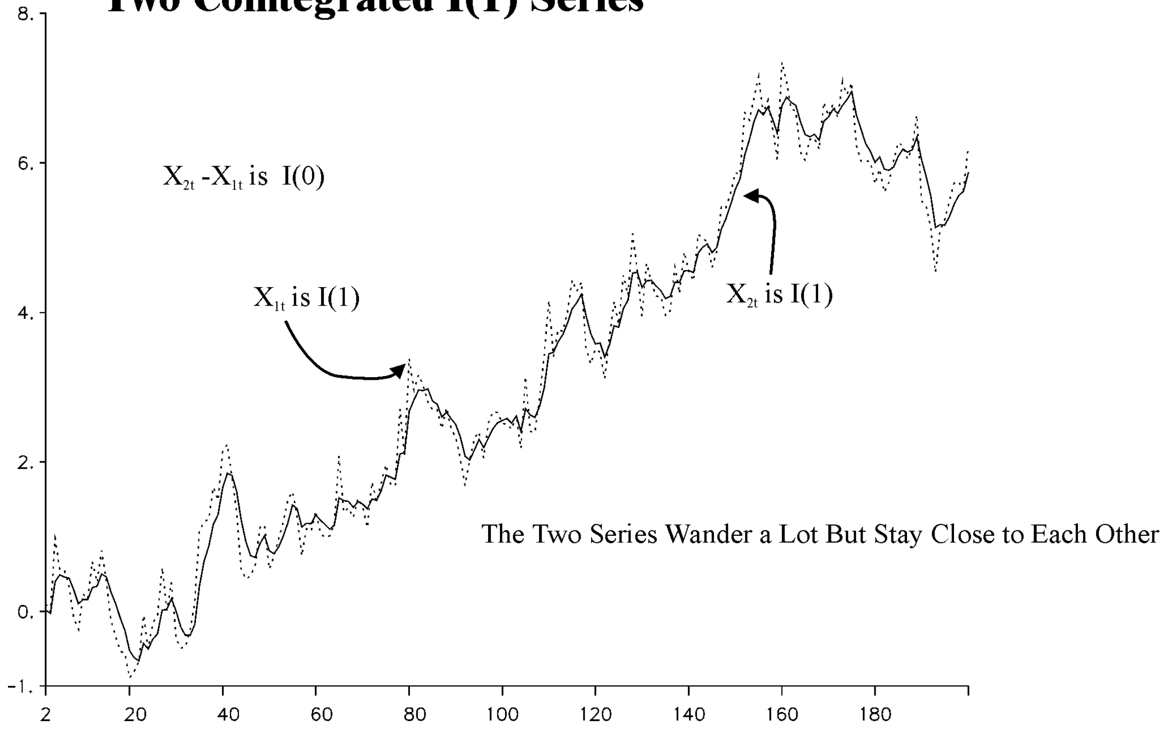
Theorem 247 *The set of cointegrating vectors that satisfy $c^T X_t \sim I(0)$ (including $c = 0$) form a linear space.*

Cointegration

I(0) and I(1) Time Series



Two Cointegrated I(1) Series



Example 248 *Given two cointegrating vectors:*

$$c_1 = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \; c_2 = \begin{bmatrix} -2 \\ 1 \\ -3 \end{bmatrix}$$

then $c = 2c_1 + 4c_2$ *given by:*

$$c = 2\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} + 4\begin{bmatrix} -2 \\ 1 \\ -3 \end{bmatrix} = \begin{bmatrix} -6 \\ 2 \\ -14 \end{bmatrix}$$

is also a cointegrating vector. Any other weights besides 2 *and* 4 *will also lead to a cointegrating vector.*

In general then if there is one cointegrating vector there will be an infinite number of cointegrating vectors. The linear space of cointegrating vectors will however contain only a finite number of linearly independent cointegrating vectors or a finite number of economically meaningful and distinct cointegrating vectors.

We might therefore rephrase our question: what determines if X_t is cointegrated, and if X_t is cointegrated, what determines the number of *linearly independent* cointegrating vectors.

Given that X_t is $I(1)$ it follows that $(1-B)X_t$ is $I(0)$ or stationary and hence has a Wold representation given by:

$$(1-B)X_t = \mu + \psi(B)a_t. \tag{6.98}$$

It turns out that whether X_t is cointegrated or not depends on the $n \times 1$ vector or growth rates μ and the $n \times n$ matrix $\psi(1)$ given by:

$$\psi(1) = I + \psi_1 + \psi_2 + \psi_3 + \cdots.$$

We have:

Theorem 249 *If the* $n \times 1$ *vector* c *is a cointegrating vector then:*

$$\begin{aligned} c^T\mu &= 0 \\ c^T\psi(1) &= 0 \end{aligned}$$

Remark 250 *The condition* $c^T\mu = 0$ *is often ignored in discussions of cointegration by simply assuming that* $\mu = 0$ *in the Wold representation. In what follows we will always assume that the condition* $c^T\mu = 0$ *is satisfied.*

The interesting object for cointegration is in fact the matrix $\psi(1)$. We have:

Theorem 251 X_t *is cointegrated if and only if* $\psi(1)$ *is a singular matrix; that is if* $\psi(1)^{-1}$ *does not exist or* $\det[\psi(1)] = 0$ *or rank* $[\psi(1)] < n$.

It turns out that it is the difference between the dimension of the matrix $\psi(1)$ and its rank which determine the number of linearly independent cointegrating vectors. In particular we have:

Theorem 252 *If rank $[\psi(1)] = n - r$ then X_t has r linearly independent cointegrating vectors $c_1, c_2, \dots c_r$ where the c_i satisfy:*

$$c_i^T \psi(1) = 0.$$

Proof. We can obtain the Beveridge-Nelson decomposition for a X_t in exactly the same way we did for a univariate time series to obtain:

$$X_t = T_t^* + Y_t^* \tag{6.99}$$

where T_t^* is an $n \times 1$ vector of $I(1)$ random walks

$$T_t^* = \mu + T_{t-1}^* + \psi(1) a_t \tag{6.100}$$

and Y_t^* is an $n \times 1$ vector which is stationary or $I(0)$ with:

$$Y_t^* = \psi^*(B) a_t \tag{6.101}$$

$$\psi^*(B) = \sum_{k=0}^{\infty} \psi_k^* B^k$$

where:

$$\psi_k^* = -\sum_{j=k+1}^{\infty} \psi_j. \tag{6.102}$$

Suppose now that X_t is cointegrated with a cointegrating vector c. By the definition of cointegration $c^T X_t$ is $I(0)$ and hence:

$$\underbrace{c^T X_t}_{I(0)} = c^T T_t^* + \underbrace{c^T Y_t^*}_{I(0)} \tag{6.103}$$

since Y_t^* is $I(0)$ from the Beveridge-Nelson decomposition. Thus from Theorem 242 if we define:

$$\tilde{T}_t^* = c^T T_t^*$$

then the only way that (6.103) can be true is if $\tilde{T}_t^*$ is $I(0)$. From (6.100) we have:

$$\tilde{T}_t^* = c^T \mu + \tilde{T}_{t-1}^* + c^T \psi(1) a_t. \tag{6.104}$$

From this we conclude that:

$$\begin{aligned} c^T \mu &= 0 \\ c^T \psi(1) &= 0 \end{aligned}$$

since if $c^T\mu \neq 0$ then $\tilde{T}_t^*$ would have a trend with slope $c^T\mu$ and hence would not be stationary or $I(0)$ while if $c^T\psi(1) \neq 0$ then $\tilde{T}_t^*$ would be a random walk with noise term: $c^T\psi(1)\,a_t$ and hence $\tilde{T}_t^*$ would be $I(1)$ and not $I(0)$ as required for cointegration. (It follows then that $\tilde{T}_t^* = \tilde{T}^*$ a constant for all t.) Recall now that the only way that $c^T\psi(1) = 0$ can be true for $c \neq 0$ is if $\psi(1)$ is a singular[4] and so it must be that $\psi(1)$ has less than full rank or:

$$rank\left[\psi(1)\right] < n.$$

The number of linearly independent cointegrating vectors r is therefore the difference between n and $rank\left[\psi(1)\right]$ and so

$$r = n - rank\left[\psi(1)\right].$$

■

If X_t has r linearly independent cointegrating vectors then we can put these into an $n \times r$ matrix C with columns given by the cointegrating vectors as:

$$C = \left[c_1, c_2, \ldots c_r\right]. \tag{6.105}$$

Therefore if Z_t is defined as $Z_t = C^T X_t$ then:

$$Z_t \equiv C^T X_t \sim I(0). \tag{6.106}$$

The matrix C is not unique however since we have:

Theorem 253 *If Γ is any non-singular $n \times n$ matrix then $\tilde{C} = C\Gamma$ gives an equivalent set of linearly independent cointegrating vectors.*

Proof. We first prove that the columns of $\tilde{C}$ are linearly independent. We have: $\tilde{C}\delta = 0$ implies that $C\Gamma\delta = 0$ which implies that $\delta = 0$ since Γ is non-singular and the columns of C are linearly independent. Thus the columns of $\tilde{C}$ are linearly independent. Furthermore:

$$\tilde{C}^T X_t = \Gamma^T C^T X_t \sim I(0)$$

since $C^T X_t$ is $I(0)$. Thus the columns of $\tilde{C}$ are cointegrating vectors. ■

6.5.3 Some Examples

Purchasing Power Parity

From (6.92) define:

$$\begin{aligned} X_{1t} &= \ln(P_t) \\ X_{2t} &= \ln\left(P_t^F\right) \\ X_{3t} &= \ln(e_t) \end{aligned} \tag{6.107}$$

[4] Otherwise if $\psi(1)^{-1}$ exists we would conclude from $c^T\psi(1) = 0$ that:

$$c^T = \psi(1)^{-1}\,0 = 0$$

which contradicts the requirement that $c \neq 0$.

so that there are $n = 3$ components in X_t. Empirically all three of these variables have strong random walk components so that it is not at all unreasonable to assume that:

$$X_t = \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} = \begin{bmatrix} \ln(P_t) \\ \ln(e_t) \\ \ln\left(P_t^F\right) \end{bmatrix} \sim I(1). \tag{6.108}$$

Now suppose that PPP held. Taking logs of both sides of (6.92) yields:

$$\ln(P_t) = \ln(e_t) + \ln\left(P_t^F\right) \tag{6.109}$$

or, we can define Z_t as the deviation from PPP as follows:

$$\begin{aligned} Z_t &= \ln(P_t) - \ln(e_t) - \ln\left(P_t^F\right) && (6.110) \\ &= X_{1t} - X_{2t} - X_{3t}. && (6.111) \end{aligned}$$

Clearly $Z_t = 0$ if and only if PPP holds.

If we now define

$$c = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} \tag{6.112}$$

then we can write the deviation from PPP as:

$$Z_t = c^T X_t. \tag{6.113}$$

If $Z_t = 0$ always, then PPP would always hold, even in the short-run. Clearly then if $Z_t = 0$ then Z_t is $I(0)$ in a trivial sense, that is Z_t is stationary. In the real world however we observe that PPP does not hold or we observe that $Z_t \neq 0$.

We might therefore wish to weaken examine a weaker form of PPP where Z_t is $I(0)$ so that PPP holds on average; that is $E[Z_t] = 0$ and that deviations from PPP never get too large. Thus if $Z_t \sim I(0)$ but we allow the possibility that $Z_t \neq 0$, equivalently that X_t is cointegrated with cointegrating vector c, then we can say that PPP holds in the long-run or on average. We can then interpret Z_t crossing the horizontal axis as indicating that the event $Z_t = 0$ occurs or that PPP holds. Thus if Z_t is $I(0)$ then from Theorem 239 PPP will hold infinitely often, although it will not hold all the time.

If on the other hand $Z_t \sim I(1)$, or X_t is not cointegrated with cointegrating vector c, then we cannot say that $E[Z_t] = 0$ since Z_t is nonstationary and hence does not have a mean. The event that $Z_t = 0$ or that PPP holds will only happen very infrequently so that even given an infinite span of time one would only see PPP holding a finite number of times. Furthermore the deviations from PPP would get arbitrarily large as time progresses. Thus if $Z_t \sim I(1)$ it would be natural to say that PPP does not hold, even in the long-run.

Note that c is not the only vector we could have used to define the deviation from PPP since by Theorem 245 $d = 3c$ or:

$$d = \begin{bmatrix} 3 \\ -3 \\ -3 \end{bmatrix} \tag{6.114}$$

also a cointegrating vector. In terms of economics however both c and d have the same economic content; that is the PPP hypothesis.

Long-Run Money Demand

Consider now an example where the cointegrating vector c might depend on unknown parameters, in particular the long-run demand for money relationship:

$$M_t^s = AY_t^\alpha R_t^\beta P_t^\gamma.$$

Taking logs of both sides we obtain:

$$\ln(M_t^s) = \ln(A) + \alpha \ln(Y_t) + \beta \ln(R_t) + \gamma \ln(P_t) \tag{6.115}$$

so that if :

$$X_t = \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \\ X_{4t} \end{bmatrix} = \begin{bmatrix} \ln(M_t^s) \\ \ln(Y_t) \\ \ln(R_t) \\ \ln(P_t) \end{bmatrix} \tag{6.116}$$

then we can define

$$Z_t = c^T X_t - \ln(A)$$

as the deviation from the long-run demand for money relationship where:

$$c = \begin{bmatrix} 1 \\ -\alpha \\ -\beta \\ -\gamma \end{bmatrix}. \tag{6.117}$$

If then $Z_t \sim I(0)$ then we can say that the long-run money demand relationship holds in the long-run and hence X_t is cointegrated with cointegrating vector c. If on the other hand $Z_t \sim I(1)$ then X_t is not cointegrated and the long-run money demand relationship does not hold.

Combining PPP and Money Demand

Consider now combining the time series in the money demand and PPP examples to form an X_t defined as:

$$X_t = \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \\ X_{4t} \\ X_{5t} \\ X_{6t} \end{bmatrix} = \begin{bmatrix} \ln(P_t) \\ \ln(e_t) \\ \ln\left(P_t^F\right) \\ \ln(M_t^s) \\ \ln(Y_t) \\ \ln(R_t) \end{bmatrix}. \tag{6.118}$$

We now have two economically meaningful long-run relationships with corresponding cointegrating vectors c_1 and c_2: PPP with c_1 and the long-run money demand relationship with c_2 as:

$$c_1 = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \; c_2 = \begin{bmatrix} -\gamma \\ 0 \\ 0 \\ 1 \\ -\alpha \\ -\beta \end{bmatrix}. \tag{6.119}$$

It does not particularly interest us as economists that say $c_3 = 4c_1 + 2c_2$, given by:

$$c_3 = 4\begin{bmatrix} 1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + 2\begin{bmatrix} -\gamma \\ 0 \\ 0 \\ 1 \\ -\alpha \\ -\beta \end{bmatrix} = \begin{bmatrix} 4-2\gamma \\ -4 \\ -4 \\ 2 \\ -2\alpha \\ -2\beta \end{bmatrix}$$

is also a cointegrating vector since c_3 merely mixes up PPP and the long-run money demand relationship in a way that is hard to interpret in an economically meaningful way. Thus we can in a sense say there are only two economically meaningful or two linearly independent cointegrating vectors c_1 and c_2 with all others being a weighted combination of these two.

Suppose we did not know c_1 and c_2 but instead estimated a model where:

$$\psi(1) = \begin{bmatrix} \frac{7}{11} & \frac{7}{22} & \frac{7}{22} & \frac{1}{22} & -\frac{1}{11} & \frac{3}{22} \\ \frac{7}{22} & \frac{29}{44} & -\frac{15}{44} & \frac{1}{44} & -\frac{1}{22} & \frac{3}{44} \\ \frac{7}{22} & -\frac{15}{44} & \frac{29}{44} & \frac{1}{44} & -\frac{1}{22} & \frac{3}{44} \\ \frac{1}{22} & \frac{1}{44} & \frac{1}{44} & \frac{41}{44} & \frac{3}{22} & -\frac{9}{44} \\ -\frac{1}{11} & -\frac{1}{22} & -\frac{1}{22} & \frac{3}{22} & \frac{8}{11} & \frac{9}{22} \\ \frac{3}{22} & \frac{3}{44} & \frac{3}{44} & -\frac{9}{44} & \frac{9}{22} & \frac{17}{44} \end{bmatrix}.$$

Using $\psi(1)$ we can determine if X_t is cointegrated and what the cointegrating vectors are.

If you calculate $\det\left[\psi(1)\right]$ (using the computer of course!) you will find that:

$$\det\left[\psi(1)\right] = 0$$

so that $\psi(1)$ is singular and hence from Theorem 251 X_t is cointegrated. Again, if you calculate the rank of $\psi(1)$ you will find (also using the computer or from the fact that $\psi(1)$ happens to be idempotent so that $rank\left[\psi(1)\right] = tr\left[\psi(1)\right] = 4$) that:

$$rank\left[\psi(1)\right] = 4.$$

Since $n = 6$ there must be $r = 6 - 4 = 2$ linearly independent cointegrating vectors from Theorem 252. Again, the computer can calculate a basis for the null space of $\psi(1)$; that is it will find two vectors c_1 and c_2 such that $c_1^T\psi(1) = 0$ and $c_2^T\psi(1) = 0$. For this particular example the computer finds:

$$c_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \\ 2 \\ -3 \end{bmatrix}, \; c_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \\ 2 \\ -3 \end{bmatrix}.$$

The first cointegrating vector c_1 is just the long-run money demand relationship. That is:

$$Z_{1t} = c_1^T X_t = \ln P_t - \ln M_t^s + 2\ln Y_t - 3\ln R_t$$

or:

$$\ln M_t^s = 2\ln Y_t - 3\ln R + \ln P_t - Z_{1t}$$

where $-Z_{1t}$ is the $I(0)$ error term and where $\alpha = 2$ is the income elasticity, $\beta = -3$ is the interest elasticity and $\gamma = 1$. It is hard to see any economic interpretation for the second cointegrating vector c_2 since:

$$Z_{2t} = c_2^T X_t = \ln e_t + \ln P_t^F - \ln M_t^s + 2\ln Y_t - 3\ln R_t.$$

However if we recall from Theorem 246 that $c = c_1 - c_2$ is also a cointegrating vector and take $Z_{3t} = (c_1 - c_2)^T X_t = Z_{1t} - Z_{2t}$ we obtain:

$$Z_{3t} = \ln P_t - \ln e_t - \ln P_t^F$$

which is the deviation from PPP .

6.5.4 The Common Trends Representation

A useful representation for a model with cointegration is:

Theorem 254 ***Common Trends Representation:*** *If X_t is cointegrated with r linearly independent cointegrating vectors then X_t can be written as:*

$$X_t = A\hat{T}_t^* + Y_t^*$$

where A is an $n \times (n-r)$ matrix such that $C^T A = 0$ where C is given in (6.105) *and $\hat{T}_t^*$ is an $(n-r) \times 1$ vector of random walks and $Y_t^* \sim I(0)$.*

If the $n \times 1$vector X_t was not cointegrated then it would contain n random walks as:

$$X_t = T_t^* + Y_t^*$$

from the Beveridge-Nelson decomposition. Thus there would be one trend or T_{it}^* for each of the n components of X_t. If X_t is cointegrated with r linearly independent cointegrating vectors, then from the common trends representation X_t can be thought of as containing only $n-r$ random walks given by the $(n-r) \times 1$ vector $\hat{T}_t^*$ rather than the full n. Thus with cointegration the n series which are $I(1)$ in X_t must share the $n-r$ random walks. Cointegration exists because this gap between n and $n-r$ allows there to be r cointegrating vectors which annihilate A; that is $C^T A = 0$.

The proof is as follows:

Proof. Given Theorem 6.103 there must be $n \times (n-r)$ matrices A and F and $(n-r) \times 1$ vector $\tilde{\mu}$ such that

$$\begin{aligned} \psi(1) &= AF^T \\ \mu &= A\mu^* \end{aligned}$$

with $C^T A = 0$. Then from (6.100)

$$\begin{aligned} \Delta T_t^* &= \mu + \psi(1) a_t \\ &= A\left(\mu^* + F^T a_t\right) \\ &= A\left(\Delta\hat{T}_t^*\right) \end{aligned}$$

where:

$$\begin{aligned} \Delta\hat{T}_t^* &= \mu^* + a_t^* \\ a_t^* &= F^T a_t. \end{aligned}$$

It then follows that:

$$T_t^* = A\hat{T}_t^*$$

so that

$$X_t = A\hat{T}_t^* + Y_t^*.$$

■

Example 255 *In the PPP long-run money demand example the common trends representation can be found by calculating a column basis for $\psi(1)$ and using these 4 vectors to form the columns of A. Here we have 6 series depending on 4 random walks as:*

$$\begin{bmatrix} \ln(P_t) \\ \ln(e_t) \\ \ln\left(P_t^F\right) \\ \ln(M_t^s) \\ \ln(Y_t) \\ \ln(R_t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} \hat{T}_{1t}^* \\ \hat{T}_{2t}^* \\ \hat{T}_{3t}^* \\ \hat{T}_{4t}^* \end{bmatrix} + \begin{bmatrix} Y_{1t}^* \\ Y_{2t}^* \\ Y_{3t}^* \\ Y_{4t}^* \\ Y_{5t}^* \\ Y_{6t}^* \end{bmatrix}.$$

where the four $\hat{T}_{it}^*$ are $I(1)$ random walks and the six Y_{it}^* are $I(0)$ or stationary.

6.5.5 The Error Correction Model

If we model X_t as DS then a natural framework to attempt an analysis of cointegration is to difference X_t and estimate an VAR(p) model as:

$$\phi(B)(1-B)X_t = \alpha + a_t. \tag{6.120}$$

Unfortunately this model rules out any possibility of cointegration! Thus we cannot use this model to study cointegration. This is an implication of Theorem 251. In particular:

Theorem 256 *If $(1-B)X_t$ follows a stationary VAR(p) process then X_t cannot be cointegrated.*

Proof. If:

$$\phi(B)(1-B)X_t = \alpha + a_t \tag{6.121}$$

then since:

$$\psi(B) = \phi(B)^{-1} \tag{6.122}$$

and:

$$\psi(1) = \phi(1)^{-1}$$

with $rank[\psi(1)] = n$ since $\phi(1)^{-1}$ exists. Therefore by Theorem 252 it follows that $r = 0$. ■

This result would seem to make the study of cointegration very difficult since so far the VAR(p) process is the only multivariate model in which estimation is tractable.

Fortunately this problem can be remedied adding an additional set of regressors to the VAR(p) model. This then is the Error Correction Model or ECM.

Theorem 257 ***Error Correction Model:*** *If X_t is $I\left(1\right)$ as*

$$\tilde{\phi}\left(B\right)X_t = \alpha + a_t$$

and X_t is cointegrated with r linearly independent cointegrating vectors, then X_t has an Error Correction Representation $\left(ECM\right)$ given by:

$$\phi\left(B\right)\Delta X_t = \alpha + DC^T X_{t-1} + a_t$$

where C and D are $n \times r$ matrices with rank r and the rows of C are composed of r linearly independent cointegrating vectors.

Proof. See the Appendix. ■

Remark 258 *If we define the matrix coefficient on X_{t-1} in the ECM as:*

$$\Gamma = DC^T$$

then Γ is uniquely defined and identifiable but D and C are not since by Theorem 253 there are many ways of representing the cointegrating vectors in C. Thus if H is any nonsingular $r \times r$ matrix then we can replace C and D by:

$$\begin{aligned}\tilde{C} &= C\left(H^{-1}\right)^T \\ \tilde{D} &= CH\end{aligned}$$

in which case $\Gamma = DC^T = \tilde{D}\tilde{C}^T$.

Remark 259 *Note that ΔX_t is stationary or $I\left(0\right)$ while X_{t-1} on the right-hand side is $I\left(1\right)$. This does not involve a contradiction because $C^T X_{t-1}$ is $I\left(0\right)$ since C is a matrix of cointegrating vectors. This becomes apparent if we define the $r \times 1$ vector Z_t as:*

$$Z_t = C^T X_t$$

or:

$$Z_t = \begin{bmatrix} Z_{1t} \\ Z_{2t} \\ Z_{3t} \\ \vdots \\ Z_{rt} \end{bmatrix} = \begin{bmatrix} c_1^T X_t \\ c_2^T X_t \\ c_3^T X_t \\ \vdots \\ c_r^T X_t \end{bmatrix} \tag{6.123}$$

so that Z_t is $I\left(0\right)$ and we can write the ECM model as:

$$\Delta X_t = \alpha + \sum_{j=1}^{r} d_j Z_{jt-1} + \sum_{j=1}^{p} \phi_j \Delta X_{t-j} + a_t \tag{6.124}$$

where d_j is the j^{th} column of D and Z_{jt-1} is the j^{th} element of Z_{t-1}.

The ECM representation is identical to a VAR(p) except that DZ_{t-1} is added to the right-hand side. Thus if you estimate a VAR(p) for ΔX_t your regression will be misspecified unless the error correction term Z_{t-1} is included as a regressor on the right-hand side.

Example 260 *In the PPP long-run money demand example the ECM would with $p = 0$ take the form:*

$$\Delta X_t = \begin{bmatrix} \Delta\ln(P_t) \\ \Delta\ln(e_t) \\ \Delta\ln\left(P_t^F\right) \\ \Delta\ln(M_t^s) \\ \Delta\ln(Y_t) \\ \Delta\ln(R_t) \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix} + \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{31} & d_{32} \\ d_{41} & d_{42} \\ d_{51} & d_{52} \\ d_{61} & d_{62} \end{bmatrix} \begin{bmatrix} Z_{1t-1} \\ Z_{2t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \\ a_{3t} \\ a_{4t} \\ a_{5t} \\ a_{6t} \end{bmatrix}$$

or:

$$\begin{aligned} \Delta\ln(P_t) &= \alpha_1 + d_{11}Z_{1t-1} + d_{12}Z_{2t-1} + a_{1t} \\ \Delta\ln(e_t) &= \alpha_2 + d_{21}Z_{1t-1} + d_{22}Z_{2t-1} + a_{2t} \\ \Delta\ln\left(P_t^F\right) &= \alpha_3 + d_{31}Z_{1t-1} + d_{32}Z_{2t-1} + a_{3t} \\ \Delta\ln(M_t^s) &= \alpha_4 + d_{41}Z_{1t-1} + d_{42}Z_{2t-1} + a_{4t} \\ \Delta\ln(Y_t) &= \alpha_5 + d_{51}Z_{1t-1} + d_{52}Z_{2t-1} + a_{5t} \\ \Delta\ln(R_t) &= \alpha_6 + d_{61}Z_{1t-1} + d_{62}Z_{2t-1} + a_{6t} \end{aligned}$$

where Z_{1t} and Z_{2t} are the deviations from PPP and the long-run money demand relationship as:

$$Z_{1t} = \ln P_t - \ln e_t - \ln P_t^F$$

$$Z_{2t} = \ln M_t^s - \alpha\ln Y_t - \beta\ln R_t - \gamma\ln P_t.$$

6.5.6 Testing for Cointegration

When the Cointegrating Vector is Known

Consider first the problem of testing for cointegration when the cointegration vector c is known. This is actually quite common in practice; for example as we saw with PPP we know that:

$$c^T = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 0 \end{bmatrix}$$

so that c does not contain any unknown parameters.

The null hypothesis and alternative hypotheses are:

$$\begin{aligned} H_o &: X_t \text{ is not cointegrated with cointegrating vector } c \text{ versus} \\ H_1 &: X_t \text{ is cointegrated with cointegrating vector } c \end{aligned}$$

which can be reformulated as:

$$\begin{aligned} H_o &: Z_t = c^T X_t \sim I(1) \\ H_1 &: Z_t = c^T X_t \sim I(0). \end{aligned}$$

For example with PPP we would calculate:

$$\begin{aligned} Z_t &= c^T X_t \\ &= \ln(P_t) - \ln(e_t) - \ln\left(P_t^F\right). \end{aligned} \tag{6.125}$$

The idea then is to determine whether Z_t, which measures the deviation from PPP, is $I(1)$, which implies that PPP does not hold even in the long-run, or if Z_t is $I(0)$ in which case we could say PPP holds in the long-run.

One very simple way of informally testing for cointegration then is simply to plot Z_t and see if it looks like an $I(1)$ process, which infrequently crosses the horizontal axis, or an $I(0)$ process which frequently crosses the horizontal axis.

More formally we can test for cointegration using an augmented Dickey-Fuller test to see if Z_t has a unit root. We therefore run the Dickey-Fuller regression:

$$\Delta Z_t = \alpha + \beta t + \gamma Z_{t-1} + \delta_1 \Delta Z_{t-1} + \delta_2 \Delta Z_{t-2} + \cdots + \delta_p \Delta Z_{t-p} + e_t \tag{6.126}$$

and test:

$$\begin{aligned} H_o &: \gamma = 0 \text{ (} X_t \text{ is not cointegrated) versus} \\ H_1 &: \gamma < 0 \text{ (} X_t \text{ is cointegrated)} \end{aligned}$$

using the t statistic, say τ_γ, for γ and the same critical values we used for the augmented Dickey-Fuller test. For example at the 5% significance level we would use $\tau_\gamma^c = -3.4$ so that if $\tau_\gamma > -3.4$ we would accept the null that X_t is not cointegrated with cointegrating vector c while if $\tau_\gamma < -3.4$ we would accept the null that X_t is cointegrated with cointegrating vector c.

When the Cointegrating Vector is Unknown

Suppose now that the cointegrating vector is unknown. For example we have seen with the money demand example that:

$$c = \begin{bmatrix} -\gamma \\ 0 \\ 0 \\ 1 \\ -\alpha \\ -\beta \end{bmatrix}$$

and so depends on the unknown the money demand parameters α, β, and γ.

Given:

$$Z_t = c^T X_t \tag{6.127}$$

suppose, without loss of generality, that the first element of c is non-zero and hence can be normalized to equal unity. We then have, adding the constant c_o if Z_t has a non-zero mean, that:

$$X_{1t} + c_2 X_{2t} + c_3 X_{3t} + \cdots + c_n X_{nt} = c_o + Z_t \tag{6.128}$$

or:

$$X_{1t} = c_o - c_2 X_{2t} - c_3 X_{3t} - \cdots - c_n X_{nt} + Z_t. \tag{6.129}$$

This can be thought of a regression with Z_t as the error term. Thus one would estimate:

$$X_{1t} = \hat{c}_o - \hat{c}_2 X_{2t} - \hat{c}_3 X_{3t} - \cdots - \hat{c}_n X_{nt} + \hat{Z}_t. \tag{6.130}$$

For example with the long-run money demand example we would run the regression:

$$\ln(M_t^s) = \alpha_o + \alpha \ln(Y_t) + \beta \ln(R_t) + \gamma \ln(P_t) + Z_t$$

and obtain the least squares residual $\hat{Z}_t$.

Now to test for cointegration perform a unit root test on least squares residual: $\hat{Z}_t$:

$$\Delta \hat{Z}_t = \alpha + \beta t + \gamma \hat{Z}_{t-1} + \delta_1 \Delta \hat{Z}_{t-1} + \delta_2 \Delta \hat{Z}_{t-2} + \cdots + \delta_p \hat{Z}_{t-p} + e_t \tag{6.131}$$

and test:

$$\begin{aligned} H_o &: \gamma = 0 \text{ (} X_t \text{ is not cointegrated) versus} \\ H_1 &: \gamma < 0 \text{ (} X_t \text{ is cointegrated)} \end{aligned}$$

using the t statistic for γ: τ_γ. Because we estimated the cointegrating vector it turns out that we need to use a different critical value than when we knew c. That is we would not use $\tau_\gamma^c = -3.4$ at the 5% level. The appropriate critical values are however available, and are provided, for example, by the CDF command in TSP.

6.5.7 The Engle-Granger Two-Step Estimation Procedure

We will now discuss the Engle-Granger two-step procedure for estimating an error correction model or ECM when $r = 1$ so that is when there is only one cointegrating vector for X_t. The case where $r > 1$ is trickier because, as we have seen, there are many different ways of representing the two or more cointegrating vectors. This is best handled with the Johansen procedure considered in the next section.

First suppose we know c (as with the PPP example). The first step then is to calculate the scalar series:

$$Z_t = c^T X_t. \tag{6.132}$$

To estimate the ECM then we merely have to add Z_{t-1} as an additional regressor and proceed in the same manner we estimated then for $i = 1, 2, \ldots n$ run the regressions:

$$\Delta X_{it} = \alpha_i + d_i Z_{t-1} + \sum_{k=1}^{p} \sum_{j=1}^{n} \phi_{ij}^{k} \Delta X_{jt-k} + a_{it}. \tag{6.133}$$

For example if there was only the PPP relationship we would calculate:

$$Z_t = \ln(P_t) - \ln(e_t) - \ln\left(P_t^f\right)$$

and run the regressions:

$$\begin{bmatrix} \Delta\ln(P_t) \\ \Delta\ln(e_t) \\ \Delta\ln\left(P_t^F\right) \\ \Delta\ln(M_t^s) \\ \Delta\ln(Y_t) \\ \Delta\ln(R_t) \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix} + \begin{bmatrix} d_{11}Z_{t-1} \\ d_{21}Z_{t-1} \\ d_{31}Z_{t-1} \\ d_{41}Z_{t-1} \\ d_{51}Z_{t-1} \\ d_{61}Z_{t-1} \end{bmatrix} + \sum_{j=1}^{p} \phi_j \begin{bmatrix} \Delta\ln(P_{t-j}) \\ \Delta\ln(e_{t-j}) \\ \Delta\ln\left(P_{t-j}^F\right) \\ \Delta\ln\left(M_{t-j}^s\right) \\ \Delta\ln(Y_{t-j}) \\ \Delta\ln(R_{t-j}) \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \\ a_{3t} \\ a_{4t} \\ a_{5t} \\ a_{6t} \end{bmatrix}$$

so that Z_{t-1} would simply be added as an extra regressor in the VAR(p) model.

Suppose now that the cointegrating vector c is unknown. The first step of the Granger-Engle two-step procedure now is to run the regression:

$$X_{1t} = c_o - c_2 X_{2t} - c_3 X_{3t} - \cdots - c_n X_{nt} + Z_t \tag{6.134}$$

in order to estimate c and obtain an estimate of $Z_t : \hat{Z}_t$, the least squares residual. It turns out estimates of coefficients are superefficient, that is the converge at a rate of T^{-1} to the true values so as far as asymptotic theory is concerned in next step, can treat $\hat{Z}_t$ as if it were Z_t. We now estimate the regressions:

$$\Delta X_{it} = \alpha_i + d_i \hat{Z}_{t-1} + \sum_{k=1}^{p} \phi_{ij} \Delta X_{jt-k} + a_{it}. \tag{6.135}$$

For example with the long-run money demand relationship as our cointegrating relationship we would run the regression:

$$\ln(M_t^s) = \alpha_o + \alpha \ln(Y_t) + \beta \ln(R_t) + \gamma \ln(P_t) + Z_t$$

and obtain:

$$\hat{Z}_t = \ln(M_t^s) - \hat{\alpha}_o - \hat{\alpha}\ln(Y_t) - \hat{\beta}\ln(R_t) - \hat{\gamma}\ln(P_t)$$

and put $\hat{Z}_{t-1}$ in the ECM regression as we did for the PPP example. Since the estimators $\hat{\alpha}, \hat{\beta}$ and $\hat{\gamma}$ are superefficient we can, for asymptotic theory purposes, treat $\hat{Z}_{t-1}$ as if it were equal Z_{t-1}, that is as if the estimated values of α, β and γ were the true values.

6.5.8 The Johansen Procedure

The Johansen procedure allows for the estimation of the ECM with any value of r and also provides a way of determining r. Let us write the ECM as:

$$\Delta X_t = \alpha + \Gamma X_{t-1} + \sum_{i=1}^{p} \phi_i \Delta X_{t-1} + a_t \tag{6.136}$$

where:

$$\Gamma = DC^T. \tag{6.137}$$

Consider then estimating (6.136) by maximum likelihood subject to the constraint that Γ be of rank r or that (6.137) holds. This is known as a reduced rank regression.

To do this perform the following regressions:

$$\begin{aligned} \Delta X_t &= b + \sum_{i=1}^{p} B_i \Delta X_{t-1} + e_{1t} \qquad (6.138) \\ X_{t-1} &= d + \sum_{i=1}^{p} D_i \Delta X_{t-1} + e_{2t} \end{aligned}$$

and construct the following $n \times n$ matrices:

$$\begin{aligned} S_{11} &= \frac{1}{T}\sum_{t=1}^{T} \hat{e}_{1t}\hat{e}_{1t}^T \qquad (6.139) \\ S_{12} &= \frac{1}{T}\sum_{t=1}^{T} \hat{e}_{1t}\hat{e}_{2t}^T \\ S_{21} &= \frac{1}{T}\sum_{t=1}^{T} \hat{e}_{2t}\hat{e}_{1t}^T = S_{12}^T \\ S_{22} &= \frac{1}{T}\sum_{t=1}^{T} \hat{e}_{2t}\hat{e}_{2t}^T. \end{aligned}$$

The positive definite matrix S_{22} can be decomposed as:

$$S_{22} = S_2 S_2^T \tag{6.140}$$

where S_2, a non-singular square matrix, can be thought of as the square root of S_{22}. Calculating S_2 is standard in most regression packages and so this is not a difficult calculation. For example we could use the Cholesky decomposition where S_2 is upper triangular, or $S_2 = C\Lambda^{\frac{1}{2}}C^{-1}$ where $S_{22} = C\Lambda C^{-1}$ is the diagonalization of S_{22} and Λ is a diagonal matrix of eigenvalues.

Now calculate:

$$S = S_2^{-1} S_{21} S_{11}^{-1} S_{12} \left(S_2^T\right)^{-1} \tag{6.141}$$

and calculate the eigenvalues λ_i for $i = 1, 2, \dots n$ (which are bounded between 0 and 1) and eigenvectors v_i for $i = 1, 2, \dots n$ of the matrix S Again calculating eigenvalues and eigenvectors is now standard in most regression packages and so this is also not a difficult calculation.

Now order the eigenvalues from the largest to the smallest as:

$$1 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0 \tag{6.142}$$

along with the corresponding eigenvectors: v_i for $i = 1, 2, \dots n$. It can then be shown that the restricted ML estimates of C and D are given by:

$$\begin{aligned} \hat{C} &= \left[\begin{array}{ccccc} \hat{c}_1 & \hat{c}_2 & \hat{c}_3 & \cdots & \hat{c}_r \end{array} \right] \\ \hat{D} &= -S_{12}\hat{C}\left(\hat{C}^T S_{22}\hat{C}\right)^{-1} \end{aligned} \tag{6.143}$$

where:

$$\hat{c}_i = S_2^T v_i \text{ for } i = 1, 2, \dots r. \tag{6.144}$$

Suppose now we wish to test:

$$\begin{aligned} H_o &: r = r_o \\ H_o &: r > r_o. \end{aligned}$$

The likelihood ratio test statistic for this null, referred to as the trace test, is given by:

$$\Lambda = -T \sum_{i=r_o+1}^{n} \ln\left(1 - \lambda_i\right). \tag{6.145}$$

Alternatively suppose that we wish to test:

$$\begin{aligned} H_o &: r = r_o \\ H_o &: r = r_o + 1 \end{aligned}$$

then the likelihood ratio test statistic for this null is given by:

$$\Lambda = -T \ln\left(1 - \lambda_{r_o+1}\right). \tag{6.146}$$

Critical values tables are available in the literature or calculated in packages such as CATS and TSP (for (6.146)). By beginning with $r_o = 0$, that is that there are no cointegrating vectors and increasing r_o until the null of $r = r_o$ is accepted, it is possible estimate the number of cointegrating vectors r.

A Numerical Example

The Johansen procedure is numerically intensive and so best left to computers for most of the calculations. We can however illustrate the nature these calculations and some of the numerical techniques you will need if you wish to do your own programming.

Suppose you have $n = 3$ time series and you find that:

$$\begin{aligned} S_{11} &= \begin{bmatrix} 0.28571 & 0.09862 & 0.14665 \\ 0.09862 & 0.32264 & 0.26366 \\ 0.14665 & 0.26366 & 0.40862 \end{bmatrix}, \ S_{22} = \begin{bmatrix} 0.21471 & 0.12554 & 0.20232 \\ 0.12554 & 0.21300 & 0.20733 \\ 0.20232 & 0.20733 & 0.29659 \end{bmatrix} \\ S_{12} &= \begin{bmatrix} 0.12917 & 0.17447 & 0.17811 \\ 0.14275 & 0.19244 & 0.22765 \\ 0.20437 & 0.16144 & 0.22248 \end{bmatrix}, \ S_{21} = \begin{bmatrix} 0.12917 & 0.14275 & 0.20437 \\ 0.17447 & 0.19244 & 0.16144 \\ 0.17811 & 0.22765 & 0.22248 \end{bmatrix}. \end{aligned}$$

In performing the Johansen procedure our first task is to decompose S_{22} as $S_{22} = S_2 S_2^T$. This can be done either with the Cholesky decomposition or by using the computer to decompose S_{22} as $S_{22} = C\Lambda C^{-1}$ where Λ is a diagonal matrix with the positive eigenvalues of S_{22} on the diagonal as:

$$\begin{aligned} S_{22} &= \begin{bmatrix} 0.21471 & 0.12554 & 0.20232 \\ 0.12554 & 0.21300 & 0.20733 \\ 0.20232 & 0.20733 & 0.29659 \end{bmatrix} \\ &= \begin{bmatrix} 0.19684 & 0.53726 & 0.2659 \\ 0.22952 & -0.49792 & 0.2684 \\ -0.32468 & -2.6256\times 10^{-2} & 0.35094 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 2.7371\times 10^{-2} & 0 & 0 \\ 0 & 8.8475\times 10^{-2} & 0 \\ 0 & 0 & .60845 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 1.0 & 1.166 & -1.6494 \\ 1.0 & -0.92678 & -0.04887 \\ 1.0 & 1.0094 & 1.3198 \end{bmatrix}. \end{aligned}$$

We then have: $S_2 = C\Lambda^{\frac{1}{2}}C^{-1}$ or:

$$\begin{aligned} S_2 &= \begin{bmatrix} 0.19684 & 0.53726 & 0.2659 \\ 0.22952 & -0.49792 & 0.2684 \\ -0.32468 & -2.6256\times 10^{-2} & 0.35094 \end{bmatrix} \\ &\quad \times \begin{bmatrix} \sqrt{2.7371\times 10^{-2}} & 0 & 0 \\ 0 & \sqrt{8.8475\times 10^{-2}} & 0 \\ 0 & 0 & \sqrt{0.60845} \end{bmatrix} \\ &\quad \times \begin{bmatrix} 1.0 & 1.166 & -1.6494 \\ 1.0 & -0.92678 & -0.04887 \\ 1.0 & 1.0094 & 1.3198 \end{bmatrix} \end{aligned}$$

so that:

$$S_2 = \begin{bmatrix} 0.39978 & 9.9228\times 10^{-2} & 0.21222 \\ 9.9228\times 10^{-2} & 0.39286 & 0.22092 \\ 0.21222 & 0.22092 & 0.45027 \end{bmatrix}.$$

Note that with this decomposition S_2 is symmetric. We now calculate S as:

$$S = S_2^{-1} S_{21} S_{11}^{-1} S_{12} \left(S_2^T\right)^{-1}$$

or:

$$S=\begin{bmatrix}0.33801 & 9.4053\times10^{-2} & 0.12041\\ 9.4048\times10^{-2} & 0.5036 & 0.35477\\ 0.12041 & 0.35477 & 0.32011\end{bmatrix}.$$

Again we ask the computer to find the three eigenvectors and eigenvalues of S: v_i and λ_i as:

$$\begin{aligned}v_1 &= \begin{bmatrix}0.4166\\ 1.0638\\ 0.84919\end{bmatrix}\leftrightarrow\lambda_1=0.82362\\ v_2 &= \begin{bmatrix}0.96083\\ -0.32185\\ -6.8155\times10^{-2}\end{bmatrix}\leftrightarrow\lambda_2=0.29796\\ v_3 &= \begin{bmatrix}9.6069\times10^{-2}\\ 0.40396\\ -0.5532\end{bmatrix}\leftrightarrow\lambda_3=0.040135.\end{aligned}$$

Note the eigenvalues are bounded between 0 and 1.

Suppose we have $T=150$ observations and we wish to test $H_o: r=0$ versus $H_1: r>0$ using the trace test the likelihood ratio test statistic would be:

$$\begin{aligned}\Lambda &= -150\left(\ln\left(1-0.82362\right)+\ln\left(1-0.29796\right)+\ln\left(1-0.040135\right)\right)\\ &= 319.48\end{aligned}$$

while if one tested $H_o: r=0$ versus $H_1: r=1$ the likelihood ratio test statistic would be:

$$\begin{aligned}\Lambda &= -150\left(\ln\left(1-0.82362\right)\right)\\ &= 260.27.\end{aligned}$$

To test $H_o: r=1$ versus $H_1: r>1$ using the trace test the likelihood ratio test statistic would be:

$$\begin{aligned}\Lambda &= -150\left(\ln\left(1-0.29796\right)+\ln\left(1-0.040135\right)\right)\\ &= 59.209\end{aligned}$$

while to test $H_o: r=1$ versus $H_1: r=2$ the likelihood ratio test statistic would be:

$$\begin{aligned}\Lambda &= -150\left(\ln\left(1-0.29796\right)\right)\\ &= 53.065.\end{aligned}$$

Critical values for these test statistics are available.

Suppose we believe that $r = 2$. To calculate the cointegrating vectors we use: $\hat{c}_i = S_2^T v_i$ for $i = 1, 2$ to obtain:

$$\begin{aligned}
\hat{c}_1 &= \begin{bmatrix} 0.39978 & 9.9226\times10^{-2} & 0.21222 \\ 9.9226\times10^{-2} & 0.39286 & 0.22092 \\ 0.21222 & 0.22092 & 0.45027 \end{bmatrix} \begin{bmatrix} 0.4166 \\ 1.0638 \\ 0.84919 \end{bmatrix} \\
&= \begin{bmatrix} 0.45232 \\ 0.64687 \\ 0.70579 \end{bmatrix} \\
\hat{c}_2 &= \begin{bmatrix} 0.39978 & 9.9226\times10^{-2} & 0.21222 \\ 9.9226\times10^{-2} & 0.39286 & 0.22092 \\ 0.21222 & 0.22092 & 0.45027 \end{bmatrix} \begin{bmatrix} 0.96083 \\ -0.32185 \\ -6.8155\times10^{-2} \end{bmatrix} \\
&= \begin{bmatrix} 0.33772 \\ -4.6158\times10^{-2} \\ 0.10212 \end{bmatrix}
\end{aligned}$$

so that

$$\hat{C} = \begin{bmatrix} 0.45232 & 0.33772 \\ 0.64687 & -4.6158\times10^{-2} \\ 0.70579 & 0.10212 \end{bmatrix}.$$

Now from

$$\hat{D} = -S_{12}\hat{C}\left(\hat{C}^T S_{22}\hat{C}\right)^{-1}$$

we have:

$$\hat{D} = \begin{bmatrix} 0.6018 & 5.5455\times10^{-2} \\ 0.70865 & 6.4549\times10^{-2} \\ 0.7171 & 8.6944\times10^{-2} \end{bmatrix}$$

so that $\hat{\Gamma}$ in the ECM model is:

$$\begin{aligned}
\hat{\Gamma} &= \hat{D}\hat{C}^T \\
&= \begin{bmatrix} 0.29093 & 0.38673 & 0.43041 \\ 0.34234 & 0.45542 & 0.50675 \\ 0.35372 & 0.45986 & 0.51500 \end{bmatrix}.
\end{aligned}$$

6.6 Appendix

6.6.1 Proof of the ECM Representation

Suppose that X_t follows a VAR(p) process in levels of the form:

$$\tilde{\phi}(B)\,X_t = \alpha + a_t. \tag{6.147}$$

If X_t is DS then $(1-B)\,X_t$ has a Wold representation:

$$(1-B)\,X_t = \mu + \psi\,(B)\,a_t. \tag{6.148}$$

As we have seen, cointegration depends on the matrix $\psi\,(1)$. Multiplying (6.147) by $(1-B)$ it follows that:

$$\tilde{\phi}\,(B)\,(1-B)\,X_t = (1-B)\,a_t \tag{6.149}$$

and substituting (6.148) in we have:

$$\tilde{\phi}\,(B)\,(\mu + \psi\,(B)\,a_t) = (1-B)\,I a_t. \tag{6.150}$$

Applying $E\,[\,]$ to both sides and using $E\,[a_t] = 0$ we conclude that:

$$\tilde{\phi}\,(1)\,\mu = 0$$

so that equating both sides we conclude that:

$$\tilde{\phi}\,(B)\,\psi\,(B) = (1-B)\,I \tag{6.151}$$

and so we have:

Theorem 261 *Given* (6.147) *and* (6.148) :

$$\begin{aligned}\tilde{\phi}\,(1)\,\mu &= 0 \\ \tilde{\phi}\,(B)\,\psi\,(B) &= (1-B)\,I.\end{aligned} \tag{6.152}$$

Proof. Setting $B = 1$ in (6.151) we arrive at:

$$\tilde{\phi}\,(1)\,\psi\,(1) = 0. \tag{6.153}$$

It then follows from Theorem 6.103 that if X_t is cointegrated then the rows of $\tilde{\phi}\,(1)$ are cointegrating vectors and hence linear combinations of C given in (6.105). We can therefore write $\tilde{\phi}\,(1)$ as:

$$-\tilde{\phi}\,(1) = DC^T \tag{6.154}$$

where D is an $n \times r$ matrix with rank r. Subtracting $-\tilde{\phi}\,(1)\,X_{t-1}$ from both sides of (6.147) we obtain:

$$\left(\tilde{\phi}\,(B) - \tilde{\phi}\,(1)\,B\right) X_t = \alpha - \tilde{\phi}\,(1)\,X_{t-1} + a_t. \tag{6.155}$$

Define:

$$\Gamma\,(B) = \tilde{\phi}\,(B) - \tilde{\phi}\,(1)\,B \tag{6.156}$$

and note that since $\Gamma\,(1) = 0$, an $n \times n$ matrix of zeros, so we can factor out a $1-B$ from $\Gamma\,(B)$ as:

$$\Gamma\,(B) = \tilde{\phi}\,(B) - \tilde{\phi}\,(1)\,B = (1-B)\,\phi\,(B). \tag{6.157}$$

Replacing $\tilde{\phi}\,(B) - \tilde{\phi}\,(1)\,B$ with $(1-B)\,\phi\,(B)$ in (6.155) we obtain:

$$\phi\,(B)\,(1-B)\,X_t = \alpha - \tilde{\phi}\,(1)\,X_{t-1} + a_t. \tag{6.158}$$

Finally replace $\tilde{\phi}\,(1)$ with the right-side of (6.154) to obtain the ECM model. ■

Chapter 7

Special Topics

7.1 The Frequency Domain

7.1.1 The Spectrum

So far we have discussed functions defined in the *time* domain. That is $\rho(k)$, ψ_k and $E_t[Y_{t+k}]$ have as arguments k, the number of time periods, which has a time dimension.

There is another way of looking at time series where the argument is a *frequency.* Suppose Y_t is a sine wave with frequency λ as:

$$Y_t = \sin(\lambda t).$$

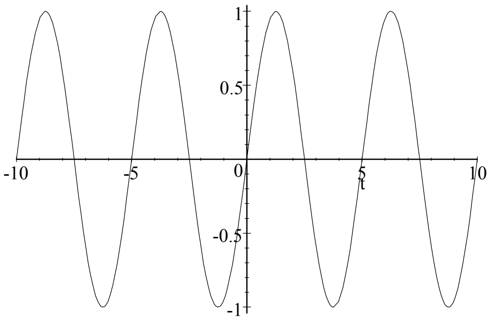
If

$$\lambda = \frac{2\pi}{P}$$

then Y_t repeats itself or has a period of P periods. For example if

$$\lambda = \frac{2\pi}{5} = 1.2566$$

then Y_t would look like:



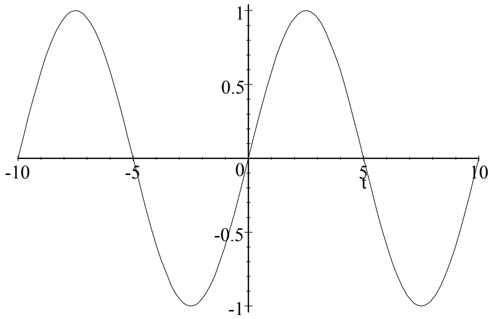
.

Plot of $\sin\left(\frac{2\pi}{5}t\right)$.

and so repeats itself every 5 periods. If we reduce the frequency so that:

$$\lambda = \frac{2\pi}{10} = 0.62832$$

then we have:



Plot of $\sin\left(\frac{2\pi}{10}t\right)$.

and Y_t now repeats itself every 10 periods. In general if

$$\lambda = \frac{2\pi}{P}$$

then Y_t repeats itself or has a period of P periods.

Economic time series do not have such regular cyclical behavior. However you can think of breaking up a time series into a number of frequencies or sine waves of different frequency. Amongst others there might for example be the seasonal frequency, a business cycle frequency. We might then be interesting in seeing how each frequency contributes to the time series in the same sense that we might want to see how much different colours or sounds contribute to a light or sound source.

The spectrum of a stationary time series: $f(\lambda)$ roughly measures the contribution of the frequency λ sine wave to Y_t. For example $\lambda = \frac{2\pi}{4}$ (which has a period of 4) would measure the importance of the seasonal frequency for a quarterly time series Y_t.

More exactly $f(\lambda)$ is the Fourier transform of the autocovariance function $\gamma(k)$ defined as:[1]

Definition 262 *The spectrum of a stationary time series: $f(\lambda)$ is (equivalently) defined as:*

$$\begin{aligned} f(\lambda) &= \frac{\gamma(0)}{2\pi} + \frac{1}{\pi}\left(\gamma(1)\cos(\lambda) + \gamma(2)\cos(2\lambda) + \gamma(3)\cos(3\lambda) + \cdots\right) \\ &= \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k)\cos(\lambda k) \\ &= \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k)e^{i\lambda k} \end{aligned}$$

for $-\pi \leq \lambda \leq \pi$.

The equivalence of the first and second lines follow from the fact that $\cos(-x) = \cos(x)$. The third line follows from Euler's Theorem:

$$e^{ix} = \cos(x) + i\sin(x)$$

and the fact that $\sin(-x) = -\sin(x)$.

Note that $f(\lambda)$ is a continuous function of the frequency λ. It can be shown that:

Theorem 263 *The spectrum is a non-negative even function:*

$$\begin{aligned} f(\lambda) &\geq 0 \\ f(-\lambda) &= f(\lambda). \end{aligned}$$

Given $\gamma(k)$ we can thus calculate the spectrum $f(\lambda)$ using the Fourier transform. It turns out that given $f(\lambda)$ we can also calculate $\gamma(k)$ using the inverse Fourier transform. Thus both $\gamma(k)$ and $f(\lambda)$ contain exactly the same amount of information, although this information is expressed in different ways. We have:

[1] To see why all three definitions are equivalent recall that:

$$\begin{aligned} \cos(-\lambda k) &= \cos(\lambda k) \\ \sin(-\lambda k) &= -\sin(\lambda k). \end{aligned}$$

Theorem 264 *The autocovariance function $\gamma(k)$ can be calculated from $f(\lambda)$ as:*

$$\begin{aligned} \gamma(k) &= \int_{-\pi}^{\pi} f(\lambda)\cos(\lambda k)\, d\lambda \qquad (7.1) \\ &= \int_{-\pi}^{\pi} f(\lambda)\, e^{-i\lambda k} d\lambda. \end{aligned}$$

By setting $k = 0$ in (7.1) it follows that

Theorem 265

$$\gamma(0) = \int_{-\pi}^{\pi} f(\lambda)\, d\lambda.$$

Remark 266 *This means that the total area under $f(\lambda)$ is the variance of Y_t given by $\gamma(0)$. Thus $f(\lambda)$ can be more precisely interpreted as the contribution of frequency λ to the total variance of Y_t.*

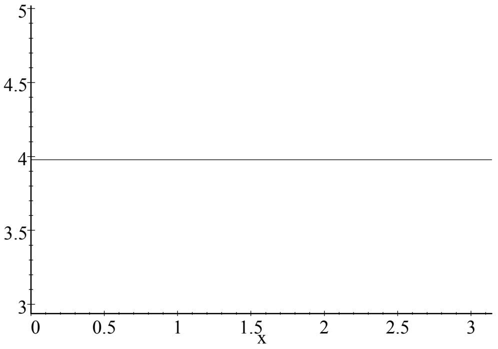
An important special case is a white noise time series so that $\gamma(k) = 0$ for $k \neq 0$. In this case the spectrum is flat and given by:

$$f(\lambda) = \frac{\gamma(0)}{2\pi} = \frac{\sigma^2}{2\pi}.$$

Thus if Y_t is white noise as:

$$Y_t = a_t \sim N[0, 25]$$

then $f(\lambda) = \frac{25}{2\pi} = 3.978$ as plotted below:



.

Spectrum of White Noise

Thus with a white noise time series all frequencies contribute equally to the variance of Y_t just as with white light all colours contribute equally to the intensity of the light.

7.1.2 The Spectrum of an ARMA(p,q)

There is a very beautiful relationship between the Wold representation and the spectrum which often allows for the easy calculation of the spectrum:

Theorem 267 *From the Wold representation: $Y_t = \psi(B)\, a_t$ the spectrum can be expressed as:*

$$\begin{aligned} f(\lambda) &= \psi\left(e^{i\lambda}\right)\psi\left(e^{-i\lambda}\right)\frac{\sigma^2}{2\pi} \\ &= \left|\psi\left(e^{i\lambda}\right)\right|^2 \frac{\sigma^2}{2\pi}. \end{aligned}$$

Remark 268 *Recall from the Beveridge-Nelson decomposition that $\psi(1)$ played a key roll. This is closely related to $f(0)$ since if $\lambda = 0$ then $e^{i\lambda}|_{\lambda=0} = e^0 = 1$ and so:*

$$f(0) = \psi(1)^2 \frac{\sigma^2}{2\pi}.$$

Remark 269 *There are multivariate analogues of the spectrum. In particular if Y_t is an $n \times 1$ multivariate vector of time series with Wold representation:*

$$Y_t = \psi(B)\, a_t$$

and $Var[a_t] = \Omega$ then the spectrum of Y_t is:

$$f(\lambda) = \psi\left(e^{-i\lambda}\right)^T \Omega \psi\left(e^{i\lambda}\right).$$

The easiest way to calculate $f(\lambda)$ is via Theorem 267. For example the AR(1) model:

$$Y_t = \phi Y_{t-1} + a_t$$

with $\psi(B) = \frac{1}{1-\phi B}$ has a spectrum:

$$\begin{aligned} f(\lambda) &= \frac{\sigma^2}{2\pi}\frac{1}{\left(1-\phi e^{i\lambda}\right)\left(1-\phi e^{-i\lambda}\right)} \\ &= \frac{\sigma^2}{2\pi}\frac{1}{\left(1+\phi^2-2\phi\cos(\lambda)\right)} \end{aligned}$$

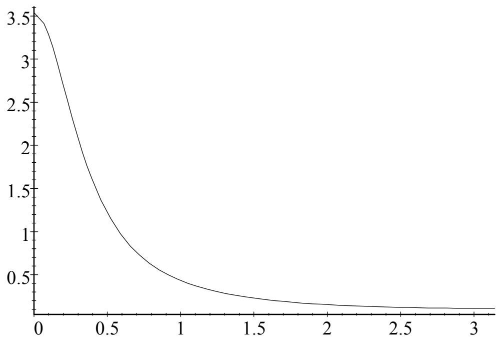
where we use the fact that:

$$\cos(\lambda) = \frac{e^{i\lambda}+e^{-i\lambda}}{2}.$$

If we set $\phi = 0.7$ and $\sigma^2 = 2$ then we obtain the spectrum:

$$\begin{aligned} f(\lambda) &= \frac{2}{2\pi}\frac{1}{\left(1-0.7e^{i\lambda}\right)\left(1-0.7e^{-i\lambda}\right)} \\ &= \frac{1}{\pi}\frac{1}{(1.49-1.4\cos(\lambda))} \end{aligned}$$

which is plotted below:



.

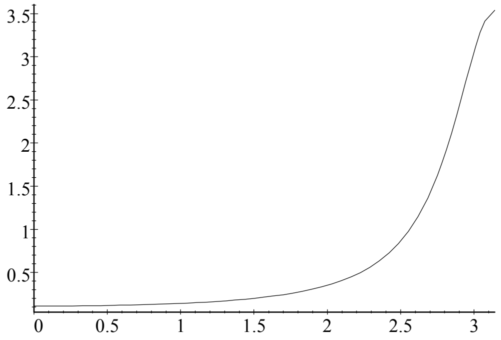
Plot of $f(\lambda)$

Note that it is the low frequencies, that is λ close to zero that contribute the most to the variance of Y_t. This reflects the fact that with $\phi > 0$ the series will exhibit persistence and hence will tend to have longer wave like patterns the greater is ϕ.

If instead set $\phi = -0.7$ with $\sigma^2 = 2$ we obtain the spectrum:

$$\begin{aligned} f(\lambda) &= \frac{2}{2\pi}\frac{1}{(1+0.7e^{i\lambda})(1+0.7e^{-i\lambda})} \\ &= \frac{1}{\pi}\frac{1}{(1.49+1.4\cos(\lambda))} \end{aligned}$$

which is plotted below:



.

Plot of $f(\lambda)$

Note that it is now the high frequencies that contribute the most to the variance of Y_t, reflecting the fact that if $Y_t > 0$ then it is more likely that $Y_{t+1} < 0$ so that Y_t tends to alternate in sign.

You might want to prove that:

Theorem 270 *The spectrum of an AR(1) is either monotonically increasing or decreasing depending on whether $\phi > 0$ or $\phi < 0$.*

From Theorem 267 the spectrum of an AR(p) is:

$$\begin{aligned} f(\lambda) &= \frac{\sigma^2}{2\pi}\psi\left(e^{i\lambda}\right)\psi\left(e^{-i\lambda}\right) \\ &= \frac{\sigma^2}{2\pi}\frac{1}{\phi\left(e^{i\lambda}\right)\phi\left(e^{-i\lambda}\right)}. \end{aligned}$$

To have a non-monotonic spectrum with a peak we need at least an AR(2). The spectrum for an AR(2) is given by:

$$\begin{aligned} f(\lambda) &= \frac{\sigma^2}{2\pi}\frac{1}{\left(1-\phi_1 e^{i\lambda}-\phi_2 e^{2i\lambda}\right)\left(1-\phi_1 e^{-i\lambda}-\phi_2 e^{-2i\lambda}\right)} \qquad (7.2) \\ &= \frac{\sigma^2}{2\pi}\frac{1}{1+\phi_1^2+\phi_2^2-2\left(\phi_1\cos(\lambda)+\phi_2\cos(2\lambda)+\phi_1\phi_2\cos(\lambda)\right)} \end{aligned}$$

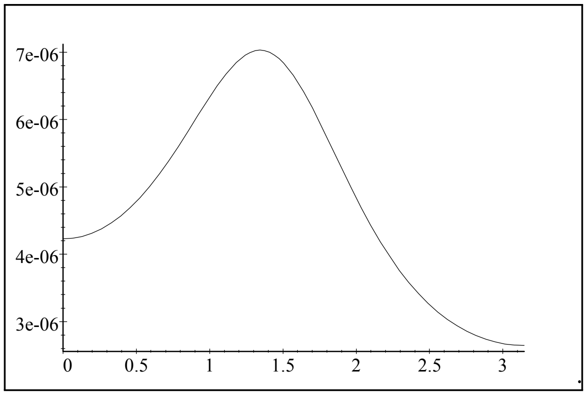
Consider for example the AR(2) in (3.135):

$$Y_t = 1.5Y_{t-1} - 0.625Y_{t-2} + a_t,\ \sigma = 0.01.$$

In this case the spectrum becomes:

$$f(\lambda) = \frac{(0.01)^2}{2\pi\left(3.6406 - 1.125\cos(\lambda) + 1.25\cos(2\lambda)\right)}$$

which is plotted below:



Plot of $f(\lambda)$

Note that $f(\lambda)$ has a peak at $\lambda = 1.3439$.[2]

[2] If you differentiate with respect to λ the first-order conditions are

$$1.125\sin(\lambda) - 2.5\sin(2\lambda) = 0.$$

From Theorem 267 the spectrum of an ARMA(p,q) is:

$$\begin{aligned} f(\lambda) &= \frac{\sigma^2}{2\pi}\frac{\theta\left(e^{i\lambda}\right)\theta\left(e^{-i\lambda}\right)}{\phi\left(e^{i\lambda}\right)\phi\left(e^{-i\lambda}\right)} \\ &= \frac{\sigma^2}{2\pi}\frac{\left|1+\theta_1 e^{i\lambda}+\theta_2 e^{i2\lambda}+\cdots+\theta_q e^{iq\lambda}\right|^2}{\left|1-\phi_1 e^{i\lambda}-\phi_2 e^{i2\lambda}-\cdots-\phi_p e^{ip\lambda}\right|^2}. \end{aligned}$$

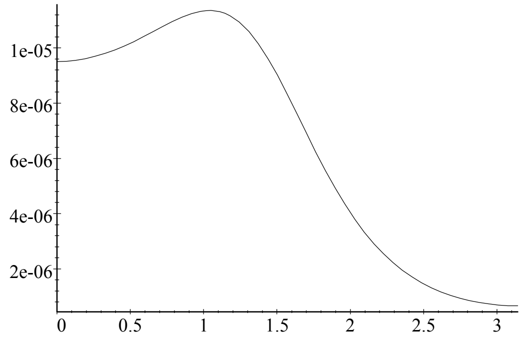
For example the ARMA(2,1) model :

$$Y_t = 1.5Y_{t-1} - 0.625Y_{t-2} + a_t + 0.5a_{t-1},\ \sigma = 0.01$$

has a spectrum:

$$f(\lambda) = \frac{(0.01)^2(1.25+\cos(\lambda))}{2\pi(3.6406 - 1.125\cos(\lambda) + 1.25\cos(2\lambda))}$$

which is plotted below:



.

Plot of $f(\lambda)$

7.1.3 Spectral Estimation

The most basic estimator of the spectrum $f(\lambda)$ is the periodogram $I(\lambda)$ defined as the Fourier transform of the estimated autocovariance function where:

Definition 271 *The periodogram $I(\lambda)$ is:*

$$I(\lambda) = \frac{1}{2\pi}\sum_{k=-(T-1)}^{T-1}\hat{\gamma}(k)\cos(\lambda k) \tag{7.3}$$

Since $\sin(2\lambda) = 2\sin(\lambda)\cos(\lambda)$ this reduces to:

$$1.125 - 5\cos(\lambda) = 0$$

or:

$$\lambda = \arccos\left(\frac{1.125}{5}\right) = 1.3439.$$

or equivalently as:

$$I(\lambda) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} Y_t e^{i\lambda t}\right|^2. \tag{7.4}$$

For cases where T is large (say $T > 1000$) it is faster to calculate $I(\lambda)$ using (7.4) and the fast Fourier transform. Otherwise (7.3) is adequate for computational purposes.

Since $I(\lambda)$ is a continuous function of λ we cannot actually calculate $I(\lambda)$ for all λ in the interval $[0,\pi]$. In fact it suffices to calculate $I(\lambda_j)$ for:[3]

$$\lambda_j = \frac{2\pi j}{T}, \text{ for } j = 0,1,2,\dots\left[\frac{T-1}{2}\right].$$

These are sometimes referred to as the fundamental frequencies. Note that for $\lambda_j = 0$ we have:

$$I(0) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} Y_t\right|^2 = \frac{T\bar{Y}^2}{2\pi}. \tag{7.5}$$

Thus if Y_t comes from either the DS or TS models where $\bar{Y} = 0$ it follows that $I(0) = 0.$

$I(\lambda_j)$ is asymptotically unbiased but the variance of $I(\lambda_j)$ does not go to zero as the sample size T increases so that $I(\lambda_j)$ is an *inconsistent* estimator of $f(\lambda)$. In fact it can be shown that:

Proposition 272 *The asymptotic distribution of the periodogram is:*

$$\frac{I(\lambda_j)}{f(\lambda_j)} \overset{a}{\sim} \chi_2^2 \text{ for } j = 1,2,\dots\left[\frac{T-1}{2}\right]$$

where $I(\lambda_j)$ and $I(\lambda_k)$ are asymptotically independent for $j \neq k$.

One way of seeing why the periodogram is not a consistent estimator is to note from (7.3) that $I(\lambda_j)$ is based on $\hat{\gamma}(k)$ for $k \leq T-1$. For $k \approx T$ there will be very few pairs of Y_t in a sample of T observations with which to estimate $\gamma(k)$ and so $\hat{\gamma}(k)$ will be a poor estimator for k large. The most extreme example is for $k = T-1$ where there is only one possible pair of observations separated by $T-1$ periods: Y_1 and Y_T so that:

$$\hat{\gamma}(T-1) = \frac{1}{T}Y_1 Y_T. \tag{7.6}$$

Clearly then $\hat{\gamma}(k)$ for $k \approx T$ will have large sample variation. This problem will remain even as $T \to \infty$ and so $I(\lambda_j)$ does not converge to the spectrum $f(\lambda)$.

Consistent estimates of $f(\lambda)$ can be obtain by giving $\hat{\gamma}(k)$ for k large a smaller weight in (7.3) using a lag window estimator:[4]

[3] $[x]$ is the integer part of x.

[4] See Priestly chapter 6

Definition 273 *The lag window estimator of $f(\lambda)$ is defined as*

$$\hat{f}(\lambda_j) = \frac{1}{2\pi} \sum_{k=-(T-1)}^{T-1} w\left(\frac{k}{M}\right) \hat{\gamma}(k) \cos(\lambda_j k) \tag{7.7}$$

where $w(\kappa)$, referred to as a lag window, has the following properties:

$$\begin{aligned} w(\kappa) &= 0 \text{ for } |\kappa| > 1 \\ w(-\kappa) &= w(\kappa) \end{aligned}$$

and where M, the width of the lag window satisfies:

$$M \to \infty$$

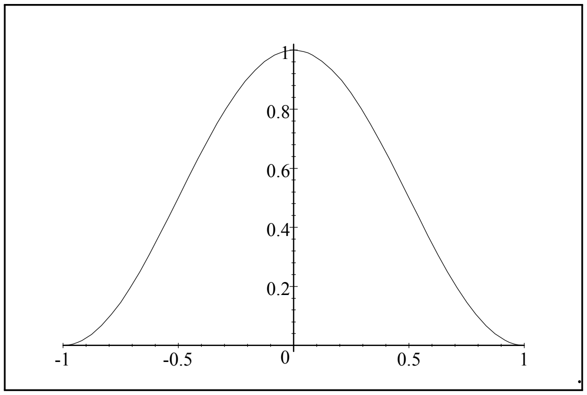
$$\frac{M}{T} \to 0 \tag{7.8}$$

as $T \to \infty$.

One popular choice, among many others for $w(\kappa)$ is the Tukey-Hanning window where:

$$w(\kappa) = \begin{cases} \frac{1}{2}(1 + \cos(\pi\kappa)), \text{if } 0 \leq |\kappa| \leq 1 \\ 0, \text{if } |\kappa| > 1 \end{cases} \tag{7.9}$$

which is plotted below:



Tukey Window: $w(\kappa)$

The fact that $w(\kappa)$ decreases as $\kappa \to 1$ insures that $\hat{\gamma}(k)$ does not affect $\hat{f}(\lambda)$ very much until T is large enough so that $\hat{\gamma}(k)$ can be reliably estimated.

With the Tukey-Hanning window it can then be shown that:

Proposition 274

$$\sqrt{\frac{T}{M}}\left(\hat{f}(\lambda_j) - f(\lambda_j)\right) \overset{a}{\sim} N\left[0, \frac{3}{4} f(\lambda_j)^2\right] \tag{7.10}$$

Note that this implies that the variance of $\hat{f}(\lambda_j)$ is approximately given by:

$$Var\left[\hat{f}(\lambda_j)\right] \approx \frac{3}{4}\frac{M}{T}f(\lambda_j)^2 \tag{7.11}$$

which does go to zero as long as (7.8) is satisfied.

Another popular choice is the Parzen window where:

$$w(\kappa) = \begin{cases} 1-6\kappa^2+6|\kappa|^3 \text{ if } 0 \leq |\kappa| \leq \frac{1}{2} \\ 2\left(1-|\kappa|^3\right) \text{ if } \frac{1}{2} < |\kappa| \leq 1 \\ 0 \text{ if } |\kappa| \geq 1. \end{cases} \tag{7.12}$$

With the Parzen window it can then be shown that:

Proposition 275

$$\sqrt{\frac{T}{M}}\left(\hat{f}(\lambda_j) - f(\lambda_j)\right) \overset{a}{\sim} N\left[0, 0.539 \times f(\lambda_j)^2\right] \tag{7.13}$$

Thus the variance of $\hat{f}(\lambda_j)$ is approximately given by:

$$Var\left[\hat{f}(\lambda_j)\right] \approx 0.539\frac{M}{T}f(\lambda_j)^2 \tag{7.14}$$

which also does go to zero as long as (7.8) is satisfied.

7.1.4 Maximum Likelihood in the Frequency Domain

It is possible to write down the likelihood for a time series model in the frequency domain. This is sometimes useful when calculating the likelihood in the time domain is very difficult; particularly with fractional differencing, which we will discuss in the next section or with exponential time series models where the Wold representation takes the form (see Bloomfield (1973)):

$$\psi(B) = \exp\left(\alpha_1 B + \alpha_1 B^2 + \cdots + \alpha_p B^p\right). \tag{7.15}$$

We have:

Theorem 276 *If Y_t has a Wold representation*

$$Y_t = \psi(B|\alpha)\, a_t$$

where α is a set of unknown parameters which determine the Wold representation: (i.e., the AR or MA parameters) then the likelihood can be approximated as:

$$l\left(\alpha, \sigma^2\right) = -\frac{T}{2}\ln\left(\sigma^2\right) - \frac{2\pi}{\sigma^2}\sum_{j=0}^{\left[\frac{T-1}{2}\right]} \frac{I(\lambda_j)}{\left|\psi\left(e^{i\lambda}|\alpha\right)\right|^2}. \tag{7.16}$$

Estimating the parameters in α thus boils down to minimizing:

$$S(\alpha) = \sum_{j=0}^{\left[\frac{T-1}{2}\right]} \frac{I(\lambda_j)}{\left|\psi\left(e^{i\lambda}|\beta\right)\right|^2} \tag{7.17}$$

over α. This can be done with a nonlinear optimization procedure, standard now in most econometric programs. Once $\hat{\alpha}$ has been calculated $\hat{\sigma}^2$ can then be calculated as:

$$\hat{\sigma}^2 = \frac{1}{2\pi T} \sum_{j=0}^{\left[\frac{T-1}{2}\right]} \frac{I(\lambda_j)}{\left|\psi\left(e^{i\lambda}|\hat{\alpha}\right)\right|^2}. \tag{7.18}$$

For example consider calculating the likelihood of an MA(1) model which has a spectrum:

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left(1+\theta^2+2\theta\cos(\lambda)\right). \tag{7.19}$$

Here $\alpha = \theta$ so that:

$$\left|\psi\left(e^{i\lambda}|\theta\right)\right|^2 = \left(1+\theta^2+2\theta\cos(\lambda)\right) \tag{7.20}$$

and the log-likelihood can be expressed as:

$$l\left(\theta,\sigma^2\right) = -\frac{T}{2}\ln\left(\sigma^2\right) - \frac{2\pi}{\sigma^2}\sum_{j=0}^{\left[\frac{T-1}{2}\right]} \frac{I(\lambda_j)}{\left(1+\theta^2+2\theta\cos(\lambda_j)\right)}. \tag{7.21}$$

A more exotic model is the first-order exponential process where:

$$\psi(B|\alpha) = e^{\alpha B} \tag{7.22}$$

so that:

$$\begin{aligned} Y_t &= e^{\alpha B} a_t \\ &= \left(1+\alpha B+\frac{\alpha^2}{2!}B^2+\frac{\alpha^3}{3!}B^3+\frac{\alpha^4}{4!}B^4+\frac{\alpha^5}{5!}B^5+\cdots\right)a_t \\ &= \sum_{k=0}^{\infty}\frac{\alpha^k}{k!}a_{t-k}. \end{aligned} \tag{7.23}$$

It would be relatively difficult to estimate α using maximum likelihood in the time domain. In the frequency domain, however, we can easily calculate the spectrum as:

$$\begin{aligned} f(\lambda|\alpha) &= \frac{\sigma^2}{2\pi}\exp\left(e^{i\lambda}\right)\exp\left(e^{-i\lambda}\right) \\ &= \frac{\sigma^2}{2\pi}\exp\left(2\alpha\cos(\lambda)\right) \end{aligned} \tag{7.24}$$

so that

$$\left|\psi\left(e^{i\lambda}|\alpha\right)\right|^2 = e^{2\alpha\cos(\lambda)} \tag{7.25}$$

and the likelihood can be calculated as:

$$l\left(\alpha,\sigma^2\right) = -\frac{T}{2}\ln\left(\sigma^2\right) - \frac{2\pi}{\sigma^2}\sum_{j=0}^{\left[\frac{T-1}{2}\right]} I\left(\lambda_j\right)e^{-2\alpha\cos(\lambda)}. \tag{7.26}$$

Either an iterative procedure or search over α can then be used to find $\hat{\alpha}$.

7.2 Fractional Differencing

So far every stationary time series model we have considered has the short-memory property. There are stationary time series models which do not have the short-memory property. One example of these are fractional differencing models.

Ordinarily we difference an integer number of times. If for example we difference Y_t d times then this can be represented as:

$$(1-B)^d\,Y_t. \tag{7.27}$$

If $d = 0$ this means that we do not difference so that:

$$(1-B)^0\,Y_t = Y_t \tag{7.28}$$

while if $d = 1$ we difference once so that:

$$(1-B)^1\,Y_t = (1-B)\,Y_t = Y_t - Y_{t-1}. \tag{7.29}$$

If $d = 2$ we difference twice and obtain:

$$\begin{aligned}(1-B)^2\,Y_t &= (1-B)\left(Y_t - Y_{t-1}\right) \\ &= Y_t - 2Y_{t-1} + Y_{t-2}\end{aligned} \tag{7.30}$$

or alternatively we could use the fact that:

$$(1-B)^2 = 1 - 2B + B^2 \tag{7.31}$$

to obtain the same result with $d = 2$.

Now it is also possible have $d = \frac{1}{2}$; that is to difference one-half times. The idea is the same as with (7.31) where we use the Taylor series of $(1-B)^{\frac{1}{2}}$ as:

$$\begin{aligned}(1-B)^{\frac{1}{2}} &= 1 - \frac{1}{2}B - \frac{1}{8}B^2 - \frac{1}{16}B^3 - \frac{5}{128}B^4 \\ &\quad -\frac{7}{256}B^5 - \frac{21}{1024}B^6 - \cdots\end{aligned} \tag{7.32}$$

In general we have:

$$\begin{aligned} (1-B)^d &= 1+(-d)\,B+\frac{1}{2}d\,(d-1)\,B^2 \\ &\quad +\left(-\frac{1}{6}d\,(d-1)\,(d-2)\right)B^3+\cdots \\ &= \sum_{k=0}^{\infty}\nu_k B^k \end{aligned} \tag{7.33}$$

where:

$$\begin{aligned} \nu_k &= (-1)^k\,\frac{d\,(d-1)\,(d-2)\cdots(d-k+1)}{k!} \\ &= (-1)^k\,\frac{\Gamma\,(k-d)}{\Gamma\,(-d)\,\Gamma\,(k+1)} \end{aligned} \tag{7.34}$$

and where $\Gamma\,(n)$ is the gamma function defined by:

$$\Gamma\,(n)=\int_0^{\infty}x^{n-1}e^{-x} \tag{7.35}$$

and which satisfies $\Gamma\,(n)=(n-1)!$ for n an integer and in general:

$$\Gamma\,(n)=(n-1)\,\Gamma\,(n-1)\,. \tag{7.36}$$

Definition 277 *We say that Y_t is a fractionally differenced time series if:*

$$(1-B)^d\,Y_t=a_t \tag{7.37}$$

or equivalently if:

$$\sum_{k=0}^{\infty}\frac{(-1)^k\,\Gamma\,(k-d)}{\Gamma\,(-d)\,\Gamma\,(k+1)}Y_{t-k}=a_t.$$

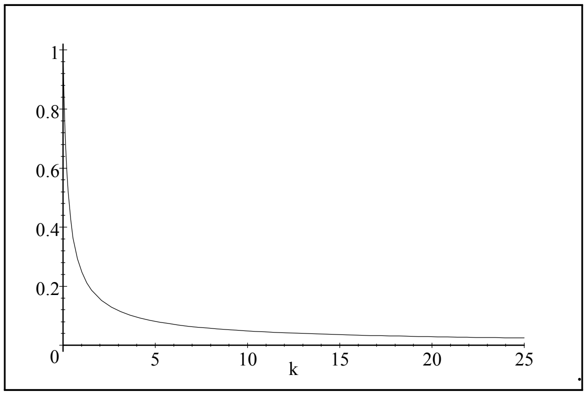
Assuming Y_t is stationary it will have a Wold representation:

$$\begin{aligned} Y_t &= (1-B)^{-d}\,a_t \\ &= \sum_{k=0}^{\infty}\psi_k a_{t-k}. \end{aligned}$$

where expanding $(1-B)^{-d}$ we have:

$$\psi_k=\frac{\Gamma\,(k+d)}{\Gamma\,(d)\,\Gamma\,(k+1)}. \tag{7.38}$$

This is plotted below for $d = \frac{1}{4}$:



ψ_k for $d = \frac{1}{4}$

From Stirling's approximation which states that for n large:

$$\Gamma(n) \approx \sqrt{2\pi n} n^n e^{-n} \tag{7.39}$$

it can be shown that for large k :

$$\psi_k = \frac{\Gamma(k+d)}{\Gamma(d)\Gamma(k+1)} \approx \frac{k^{d-1}}{\Gamma(d)}. \tag{7.40}$$

Thus ψ_k decays at a hyperbolic rate $O\left(k^{d-1}\right)$ rather than exponentially as $O\left(\tau^k\right)$ for short-memory models.

Example 278 *When $d = \frac{1}{4}$ we have:*

$$\psi_k \approx \frac{k^{-\frac{3}{4}}}{\Gamma\left(\frac{1}{4}\right)} = 0.27582 \times k^{-\frac{3}{4}}$$

and so $\psi_k = O\left(k^{-\frac{3}{4}}\right)$.

Let us now address the question of stationarity. If $d = 0$ then Y_t is white noise and hence stationary while if $d = 1$ Y_t follows a random walk and hence is not stationary. It would appear therefore that somewhere between $d = 0$ and $d = 1$ a fractionally differenced process becomes nonstationary. This transition point in fact occurs at $d = \frac{1}{2}$. We have:

Theorem 279 *Y_t in (7.37) is stationary and invertible if:*

$$-\frac{1}{2} < d < \frac{1}{2}.$$

Proof. (Informal) Stationarity requires $Var\left[Y_t\right] < \infty$. From (7.40) we have:

$$Var\left[Y_t\right] \approx \frac{\sigma^2}{\Gamma\left(d\right)^2} \sum_{k=1}^{\infty} k^{2(d-1)} < \infty.$$

In order for the sum to be finite we require the exponent on k to be less than -1 and so:

$$2\left(d-1\right) < -1$$

which implies that:

$$d < \frac{1}{2}.$$

To show $d > -\frac{1}{2}$ for invertibility requires similar ideas. ■

It can be shown that the variance $\gamma\left(0\right)$ and the autocorrelation function $\rho\left(k\right)$ for Y_t in (7.37) are given by:

$$\begin{aligned} \gamma\left(0\right) &= \sigma^2 \frac{\Gamma\left(1-2d\right)}{\Gamma\left(1-d\right)^2} \\ \rho\left(k\right) &= \frac{\Gamma\left(1-d\right)\Gamma\left(k+d\right)}{\Gamma\left(d\right)\Gamma\left(k+1-d\right)}. \end{aligned}$$

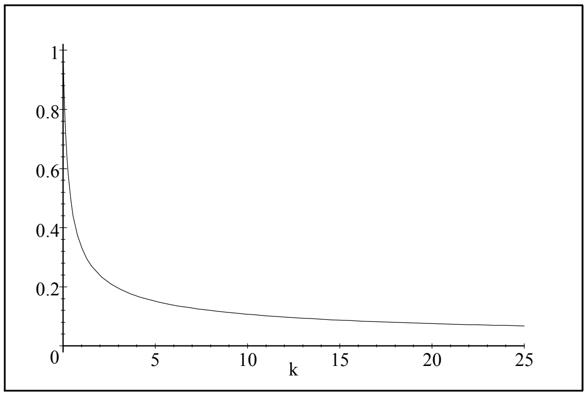
For example if $d = \frac{1}{4}$ and $\sigma^2 = 1$ we have:

$$\gamma\left(0\right) = \frac{\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{3}{4}\right)^2} = 2.6114$$

and

$$\rho\left(k\right) = \frac{\Gamma\left(\frac{3}{4}\right)\Gamma\left(k+\frac{1}{4}\right)}{\Gamma\left(\frac{1}{4}\right)\Gamma\left(k+3\right)}$$

and where $\rho\left(k\right)$ is plotted below:



Plot of $\rho\left(k\right)$, $d = \frac{1}{4}$

Note the slow rate of convergence to zero of $\rho\left(k\right)$. This reflects the fact that:

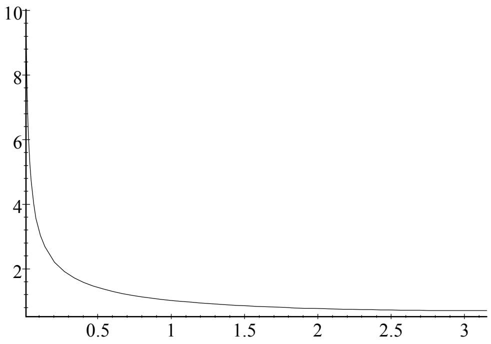
$$\rho\left(k\right)=\frac{\Gamma\left(1-d\right)\Gamma\left(k+d\right)}{\Gamma\left(d\right)\Gamma\left(k+1-d\right)}\approx\frac{\Gamma\left(1-d\right)}{\Gamma\left(d\right)}k^{2d-1}.$$

As long as the stationarity condition $d<\frac{1}{2}$ is satisfied $\rho\left(k\right)$ decays hyperbolically as $O\left(k^{2d-1}\right)$ and so $\rho\left(k\right)$ also has the long-memory property.

The spectrum of Y_t is given by:

$$\begin{aligned} f\left(\lambda\right) &= \frac{\sigma^2}{2\pi}\left(1-e^{i\lambda}\right)^{-d}\left(1-e^{-i\lambda}\right)^{-d} \\ &= \frac{\sigma^2}{2\pi}\left(2\left(1-\cos\left(\lambda\right)\right)\right)^{-d}. \end{aligned} \tag{7.41}$$

A plot of $f\left(\lambda\right)$ with $\sigma^2=2\pi$ and $d=\frac{1}{4}$ is given below:



.

$f\left(\lambda\right)$ with $d=\frac{1}{4}$

Note how $f\left(\lambda\right)\to\infty$ as $\lambda\to0$, so that the series is dominated by the low frequencies. The fact that $f\left(0\right)=\infty$ in turn implies that the autocovariances of the series sum to infinity. We have:

Theorem 280 *If* $d>0$ *then* $f\left(0\right)=\infty$ *and*

$$\sum_{k=1}^{\infty}\gamma\left(k\right)=\infty.$$

Proof. That $f\left(0\right)=\infty$ follows from (7.41). Using the definition of the spectrum and the fact that $\cos\left(0\right)=1$ we obtain:

$$\begin{aligned} f\left(0\right) &= \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma\left(k\right) \\ &= \frac{\gamma\left(0\right)}{2\pi}+\frac{1}{\pi}\sum_{k=1}^{\infty}\gamma\left(k\right)=\infty. \end{aligned}$$

■

Finally, we can add ARMA(p,q) dynamics to a fractionally differenced model to obtain a fractionally integrated ARMA or ARFIMA(p,d,q) model defined as:

Definition 281 *We say that Y_t is a ARFIMA(p,d,q) if:*

$$\phi\left(B\right)\left(1-B\right)^d Y_t = \theta\left(B\right)a_t. \tag{7.42}$$

It can be quite difficult to calculate the likelihood in the time domain. Generally the easiest way is to work with the frequency domain where the log-likelihood for an ARFIMA(p,d,q) is from (7.16):

$$l\left(\phi,\theta,d,\sigma^2\right)-\frac{T}{2}\ln\left(\sigma^2\right)-\frac{2\pi}{\sigma^2}\sum_{j=0}^{\left[\frac{T-1}{2}\right]}\frac{I\left(\lambda_j\right)\left|\theta\left(e^{\iota\lambda_j}\right)\right|^2}{\left(2\left(1-\cos\left(\lambda_j\right)\right)\right)^d\left|\phi\left(e^{\iota\lambda_j}\right)\right|^2}.$$

7.3 Nonlinearity and Nonnormality

7.3.1 A General Class of Models

Until now we have dealt only with linear time series models with normally distributed errors. What is one to do if the data do not seem consistent with either linearity or normality?

Consider a time series model which takes the following form:

$$Y_t = \mu_t\left(\phi\right)+\sqrt{h_t\left(\alpha\right)}Z_t. \tag{7.43}$$

where Z_t is *i.i.d.* with density:

$$p\left(z \mid v\right) \tag{7.44}$$

where v is a vector of parameters.

In this framework $\mu_t\left(\phi\right)$ is the conditional mean of Y_t at time based on the information at time $t-1$ while $h_t\left(\phi\right)$ is the conditional variance so that:

$$\begin{aligned} E_{t-1}\left[Y_t\right] &= \mu_t\left(\phi\right) \\ Var_{t-1}\left[Y_t\right] &= h_t\left(\alpha\right). \end{aligned} \tag{7.45}$$

Every univariate time series model we have considered so far can be considered as a special case of (7.43). This class of models however includes many more types of models. For example it would include an AR(p) with an error term that has a t distribution instead of the normal that we have assumed so far. It also includes ARCH and GARCH models as well as threshold time series models. It is also possible to extend this class of models to multivariate time series models.

If you have a model that is of the form (7.43) then it is possible to estimate it by maximum likelihood. The key result is that the log-likelihood can be shown to take the form:

$$l(\phi,\alpha,v)=-\frac{1}{2}\sum_{t=1}^{T}\ln\left(h_t(\alpha)\right)+\sum_{t=1}^{T}\ln\left(p\left(\frac{Y_t-\mu_t(\phi)}{h_t(\alpha)}\mid v\right)\right). \tag{7.46}$$

There are now generally available computer routines, for example in TSP, $GAUSS$, $RATS$ and $SHAZAM$ which will find the maximum likelihood estimates for you if you can program the log-likelihood. Thus to estimate the parameters ϕ,α and v in any of these models then you merely give such a program the above likelihood and the computer does the rest.

In what follows we will look at some very useful models that can be estimated using this approach.

7.3.2 Families of Densities

The normal distribution is not always consistent with economic data. Instead we might want to have densities that are skewed or have thicker tails than the normal. In what follows we will discuss two families of distributions which allow for thicker tails than the normal.

The density

$$p(z\mid v) \tag{7.47}$$

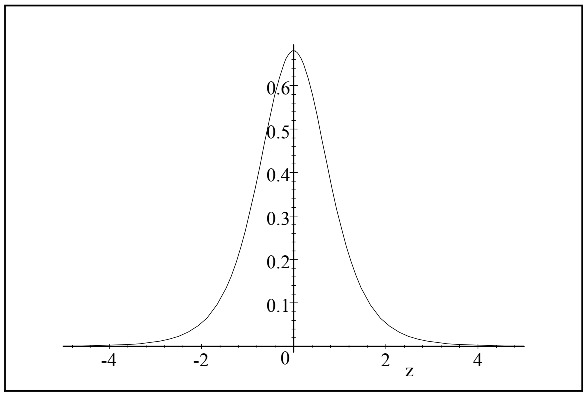
depends on a set of parameters v. If Z_t has a standard normal distribution then there are no parameters and we can write:

$$p(z)=\frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}. \tag{7.48}$$

We can allow for thicker tails (i.e., a higher probability of outliers) by assuming that Z_t has a student's t distribution with v degrees of freedom (normalized so that $Var\left[Z_t\right]=1$) where:

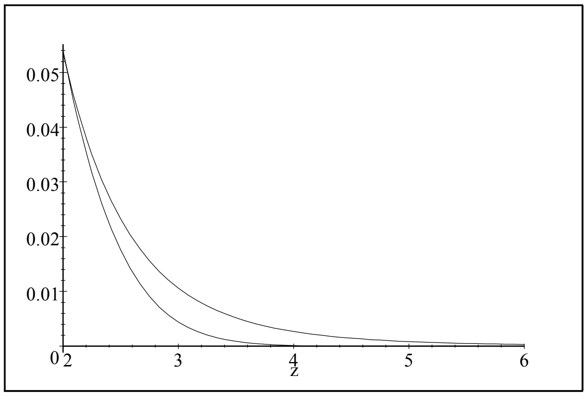
$$p(z\mid v)=\frac{\Gamma\left(\frac{v}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{v+1}{2}\right)\sqrt{v-2}}\left(1+\frac{z^2}{v-2}\right)^{-\frac{v+1}{2}}. \tag{7.49}$$

Using this framework then we can estimate v, the degrees of freedom and allow the data to decide if the normal distribution is appropriate or not since when $v=\infty$ this density reduces to the normal distribution. This density is plotted below with $v = 7$:



t distribution: $v = 7$ and $v = \infty$.

The t distribution looks like the normal distribution but has much thicker tails and hence a higher probability of outliers. This is illustrated below where the tails of the t distribution with $v = 7$ and the normal distribution are plotted for $z > 2$:



Tails of t and Normal distribution

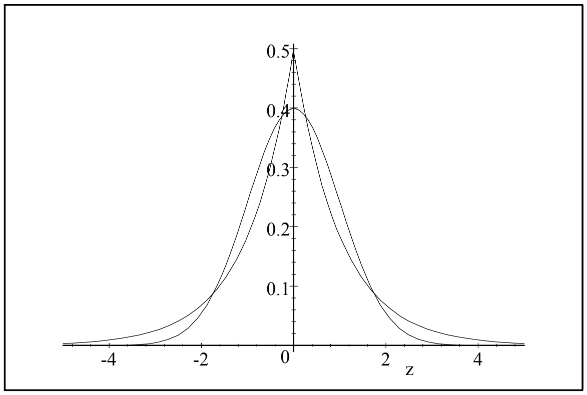
Another class of densities which have thicker tails than the normal and are often found in practice are those of the generalized exponential distribution:

$$p(z \mid v) = \frac{e^{-\frac{1}{v}|z|^{v}}}{2v^{\left(1-\frac{1}{v}\right)}\Gamma\left(\frac{1}{v}\right)}. \tag{7.50}$$

This reduces to the standard normal distribution when $v = 2$ and the Laplace distribution

$$p(z \mid 1) = \frac{e^{-|z|}}{2}$$

when $v = 1$. The normal and Laplace densities are plotted below:



Generalized Exponential: $v = 1, 2$.

7.3.3 Regime Switching Models

This class of models includes the familiar AR(p) model when we define:

$$\begin{aligned} \mu_t(\phi) &= \sum_{j=1}^{p} \phi_j Y_{t-j}, \qquad (7.51) \\ h_t(\alpha) &= \sigma^2 \\ p(z \mid v) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \end{aligned}$$

in which case we have:

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + a_t \qquad (7.52)$$

where

$$a_t = \sigma Z_t \sim N\left[0, \sigma^2\right], \qquad (7.53)$$

the ordinary AR(p) process.

By modifying $p(z \mid v)$ we could, say to be of the form (7.49) or (7.50), we obtain an AR(p) model where a_t is non-normal, say a Student's t distribution or generalized exponential distribution.

An interesting class of models is one which allows for changes in regime. An AR(p) model restricts the dynamics of the economy to be the same in all periods. We might, however, believe that the dynamics of the economy are different say in a recession than in an expansion. A common observation is that recessions are short with a dramatic decline while expansions last longer with more gentle changes.

This could be captured by supposing that Y_t is an AR(1) but with an AR parameter of ϕ_1 when $Y_t > 0$ and ϕ_2 when $Y_t < 0$ so that:

$$\mu_t(\phi_1, \phi_2) = \begin{cases} \phi_1 Y_{t-1}, \text{if } Y_{t-1} > 0 \\ \phi_2 Y_{t-1}, \text{if } Y_{t-1} \leq 0. \end{cases} \tag{7.54}$$

More generally we could have:

$$\mu_t(\phi_1, \phi_2) = \begin{cases} \sum_{i=1}^{p} \phi_i^1 Y_{t-i}, \text{if } Y_{t-d} > c \\ \sum_{i=1}^{p} \phi_i^2 Y_{t-i}, \text{if } Y_{t-d} \leq c \end{cases}$$

so that the AR parameter changes from ϕ_i^2 to ϕ_i^1 when Y_{t-d} crosses the threshold given by c. These are called threshold autoregressive processes or TAR processes. It is also possible to smooth the change in regime so that instead of an abrupt change as Y_{t-d} crosses c there is a smooth change over from the two regimes. Such models are called STAR or smooth transition threshold autoregressive processes.

TAR models have the property that the condition which determines the regime is observable; that is if:

$$S_t = \begin{cases} 1 \text{ if } Y_{t-d} > c \\ 0 \text{ if } Y_{t-d} \leq c \end{cases}$$

then at time t we observe S_t which determines which regime we are in. We can write μ_t as:

$$\mu_t(\phi_1, \phi_2) = S_t \sum_{i=1}^{p} \phi_i^1 Y_{t-i} + (1 - S_t) \sum_{i=1}^{p} \phi_i^2 Y_{t-i}.$$

Hamilton (1989) discusses a class of regime switching models, which has subsequently become very popular, where S_t is an unobserved process determined by a Markov chain with transition matrix:

$$\begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

where:

$$p_{ij} = \Pr[S_t = j | S_{t-1} = i]$$

where $i, j = 0, 1$. These models are much harder to estimate than TAR models but it is now feasible. Useful techniques include the EM algorithm and Gibbs sampling.

7.3.4 Changing Volatility

Many financial series are subject to dramatic changes in their volatility. There are a number of models which attempt to capture this behavior and which are included in this general framework.

The earliest is where Y_t follows a q^{th} order Autoregressive Conditional Heteroskedasticity or ARCH(q) process. In this case we set:

$$\begin{aligned} \mu_t(\phi) &= 0 \\ h_t(\alpha) &= \alpha_o + \alpha_1 Y_{t-1}^2 + \alpha_2 Y_{t-2}^2 + \cdots + \alpha_q Y_{t-q}^2. \end{aligned} \tag{7.55}$$

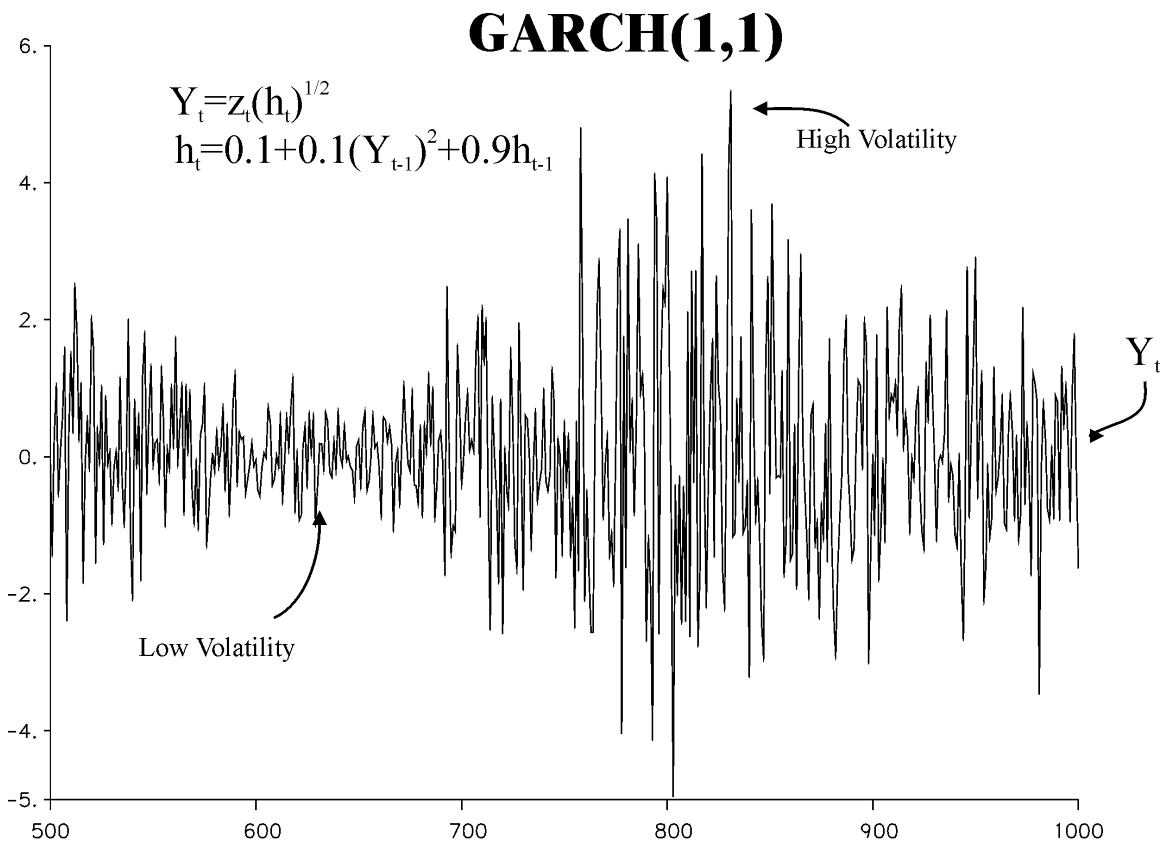
A very successful extension of the ARCH(q) model is a generalized ARCH or GARCH. The GARCH(1,1) is given by:

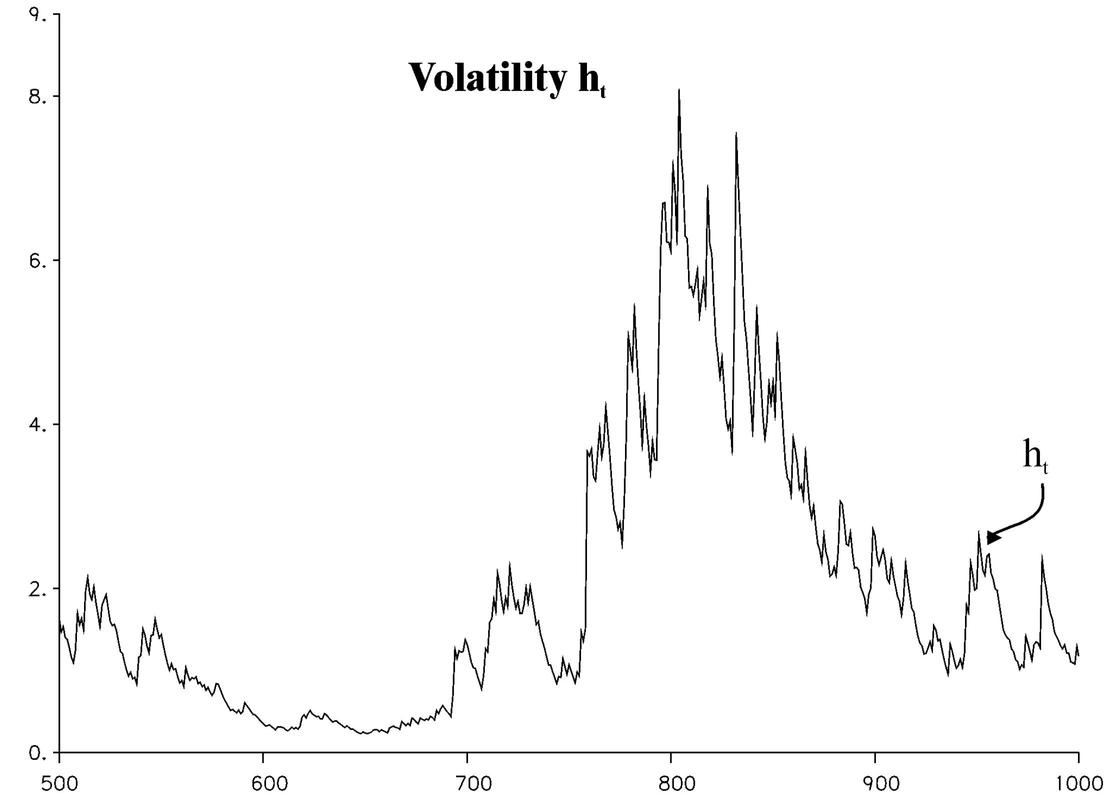
$$h_t(\alpha_o, \alpha_1, \alpha_2) = \alpha_o + \alpha_1 Y_{t-1}^2 + \alpha_2 h_{t-1}(\alpha_o, \alpha_1, \beta_1). \tag{7.56}$$

This can be extended to the GARCH(p,1) process as:

$$\begin{aligned} h_t(\alpha, \beta) &= \alpha_o + \alpha_1 Y_{t-1}^2 + \alpha_2 Y_{t-2}^2 + \cdots + \alpha_q Y_{t-q}^2 \\ &\quad + \beta_1 h_{t-1}(\alpha, \beta) + \beta_2 h_{t-2}(\alpha, \beta) + \cdots + \beta_p h_{t-p}(\alpha, \beta). \end{aligned}$$

Many other functional forms for $h_t(\alpha, \beta)$, such as EGARCH fall into this framework. Many of these models combine non-normal densities such as the t density in (7.49) or the generalized exponential distribution in (7.50).


GARCH(1,1)
$Y_t=z_t(h_t)^{1/2}$
$h_t=0.1+0.1(Y_{t-1})^2+0.9h_{t-1}$
High Volatility
Low Volatility
Y_t
6.
4.
2.
0.
-2.
-4.
-5.
500
600
700
800
900
1000


Volatility h_t
h_t
9.
8.
6.
4.
2.
0.
500
600
700
800
900
1000

References

Beveridge S. and C. Nelson (1981), “A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the Business Cycle,” *Journal of Monetary Economics* 7, 151-174.
-The Beveridge-Nelson decomposition.

Bloomfield, P. (1973) “An Exponential Model for the Spectrum of a Scalar Time Series,” *Biometrika*, 60,2, 217-226.
-An interesting paper and where the exponential time series models discussed in the text come from.

Box, G. and G. Jenkins (1976) *Time Series Analysis: Forecasting and Control*, Holden-Day Revised Edition
-Where in some sense time series analysis began and still a very useful reference.

Campbell, J. and N.G. Mankiw (1987) “Permanent and Transitory Components in Macroeconomic Fluctuations,” *American Economic Review*, 111-117.
-Closely related to the Beveridge-Nelson decomposition

Cochrane J. (1988) “How Big is the Random Walk in GNP”, *Journal of Political Economy*, 96, 893-920.
-Nice paper and closely related to the Beveridge-Nelson decomposition.

Engle, R. (1982), “Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation,” *Econometrica*, 50, 987-1008.
-The first paper on ARCH. One of the great ideas in time series analysis.

Engle, R. and T. Bollerslev, (1986) “Modelling the Persistence of Conditional Variances,” Econometric *Reviews*, 5(1), 1-50.
-Nice paper showing the power of the GARCH(1,1) model.

Engle R. and C. Granger (1987) “Co-integration and Error Correction: Representation, Estimation and Testing,” *Econometrica* 55, 251-276.
-The first paper on cointegration.

Engle R. and S. Kozicki (1993) “Testing for Common Features,” *Journal of Business and Economic Statistics*, October 1993.

Fuller, W.(1976) *Introduction to Statistical Time Series,.* Wiley.
-Advanced. Chapter 8 is good for tests for unit roots.

Geweke J. and S. Porter-Hudak (1983) “The Estimation and Application of Long Memory Time Series Models,” Journal *of Time Series Analysis*, 4, 221-238.
-Provides a method for estimating d in fractional differencing models.

Granger, C. and R. Joyeux (1980) “An Introduction to Long-Memory Time Series Models and Fractional Differencing”, *Journal of Time Series Analysis* 1, 15-29.
-Along with Hosking the first paper on fractional differencing.

Granger, C. and T. Terasvirta (1993) *Modelling Nonlinear Economic Relations, Oxford.*
-Nice discussion of nonlinear time series models, in particular regime switching models

Granger C. (1969) “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods,” *Econometrica* July 1969, 424-38.
-The original paper on Granger causality.

Hamilton, J. (1989) “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle,” *Econometrica*, 57, 357-384.
-An important paper which began a large regime switching literature.

Hamilton, J.(1994) *Time Series Analysis,* Princeton University Press.
-A good and up-to-date reference

Harvey, A. (1981) *Time Series Models,* Philip Allan.

Harvey, A. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter,* Cambridge University Press.
-Good place to start with the Kalman filter.

Hosking, J. (1981) “Fractional Differencing,” *Biometrika* 68, 165-176.
-One of the first papers on fractional differencing and long-memory.

Johansen, S. (1988), "Statistical Analysis of Cointegrating Vectors," *Journal of Economic Dynamics and Control*, 12, 231-254.

-Where the Johansen procedure was first proposed.

Johansen, S. and K. Juselius (1990) "Maximum Likelihood Estimation and Inference on Cointegration-With Applications to the Demand for Money," *Oxford Bulletin of Economics and Statistics*, 52, 169-210.

Judge G., W. Griffiths, R. Carter Hill, H. Lutkepohl. and T. Lee (1985) *The Theory and Practice of Econometrics*, second edition, Wiley.

-A good place to begin a literature search or to get a basic introduction to a subject

Lutkepohl, H. (1991) *Introduction to Multiple Time Series Analysis*, Springer-Verlag

-Everything you wanted to know (and more) about multiple time series models.

Mills, T., (1990) *Time Series Techniques For Economists.* . Cambridge University Press.

Mills, T. (1993) *The Econometric Modelling of Financial Time Series.* Cambridge University Press.

Nelson, C. and C. Plosser (1982) "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*,10,139-162.

-Influential paper that convinced many people that the economic world is DS and not TS.

Perron, P. (1988) "Trends and Random Walks in Macroeconomic Time Series," *Journal of Economics Dynamics and Control*.

-A Nice Survey of Unit Testing

Priestly, M.(1981) *Spectral Analysis and Time Series,* Academic Press.

-A good general reference. Advanced, but derivations are not cluttered with technical details so its often surprisingly easy to follow.

Sampson, M. (1991) "The Effect of Parameter Uncertainty on Forecast Variances and Confidence Intervals for Unit Root and Trend Stationary Time-Series Models," *Journal of Applied Econometrics*, 6, 67-76.

-Check it out!

Sims C. (1972) “Money, Income, and Causality,” *American Economic Review,* 62 ,4.

-An Early Application of Granger causality.

Terasvirta, T. and H. Anderson (1992) “Characterizing nonlinearities in business cycles using smooth transition autoregressive models,” Journal *of Applied Econometrics* 7, S119-S136.

-Application of the regime switching STAR model. Very interesting!