Stationary Stochastic Time Series Models

When modeling time series it is useful to regard an observed time series, $(x_1, x_2, ..., x_n)$, as the realisation of a stochastic process. In general a stochastic process can be described by an n - dimensional probability distribution $p(x_1, x_2, ..., x_n)$ so that the relationship between a realisation and a stochastic process is analogous to that between the sample and population in classical statistics.

Specifying the complete form of the probability distribution will in general be too ambitious so we usually content ourselves with the first and second moments, that is, (i) the n means, (ii) the n variances and (iii) the $n(n-1)/2$ covariances.

(i) $E(x_1), E(x_2), ..., E(x_n)$

(ii) $V(x_1), V(x_2), ..., V(x_n)$

(iii) $Cov(x_i, x_j)$, $i < j$.

If we could assume joint normality of the distribution, these set of conditions would then completely characterise the properties of the stochastic process. Even if this were the case, it will be impossible to infer all values of the first and second moments from just one realisation of the process, since there are only n observations but n (means)+ n (variances) + $n(n-1)/2$ (covariances) unknown parameters.

Further simplifying assumptions must be made to reduce the number of unknown parameters to manageable proportions.

Stationarity

A stochastic process is said to be strictly stationary if its properties are unaffected by a change in the time origin, that is

$$p(x_1, x_2, ..., x_n) = p(x_{1+l}, x_{2+l}, ..., x_{n+l}).$$

A stochastic process is said to be *weak stationary* if the first and second moments exist and do not depend on time.

$$E(x_1) = E(x_2) = ... = E(x_t) = \mu \tag{1}$$

$$V(x_1) = V(x_2) = ... = V(x_t) = \sigma^2 \tag{2}$$

$$Cov(x_t, x_{t-k}) = Cov(x_{t+l}, x_{t-k+l}) = \gamma_k \tag{3}$$

Condition (3) states that the covariances are functions only of the lag k, and not of time. These are usually called **autocovariances**.

From conditions (2) and (3) we can easily derive that the **autocorrelations**, denoted as ρ_k also only depend on the lag.

$$\rho_k = \frac{Cov(x_1, x_2)}{\sqrt[2]{V(x_1)V(x_2)}} = \frac{\gamma_k}{\sigma^2} = \frac{\gamma_k}{\gamma_o} \tag{4}$$

The **autocorrelations** considered as a function of k are referred to as the autocorrelation function, **ACF**, or sometimes the correlogram. Note that since

$$\gamma_k = Cov(x_t, x_{t-k}) = Cov(x_{t-k}, x_t) = Cov(x_t, x_{t+k}) = \gamma_{-k}$$

it follows that $\gamma_k = \gamma_{-k}$, and only the positive half of the acf is usually given.

The Wold decomposition theorem

Every weakly stationary, purely non-deterministic, stochastic process $(x_t - \mu)$ can be written as a linear combination of uncorrelated random variables. (by purely non-deterministic we mean that any linear deterministic components have already been subtracted from x_t).

This representation is given by

$$\begin{aligned} (x_t - \mu) &= \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ... \\ &= \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} \quad \text{where } \theta_0 = 1 \end{aligned} \tag{5}$$

The sequence of random variables $(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, ...)$ are assumed to be uncorrelated and identically distributed with zero mean and constant variance (a white-noise process), that is

$$\begin{gathered} E(\varepsilon_t) = 0 \\ V(\varepsilon_t) = \sigma^2 \\ Cov(\varepsilon_t, \varepsilon_{t-k}) = 0 \text{ for all } k. \end{gathered}$$

Using equation (5) we can see that;

The mean of the process described in equation (5) is

$$E(x_t) = \mu, \tag{6}$$

The Variance is

$$\begin{aligned} \gamma_o &= E(x_t - \mu)^2 \\ &= E(\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ...)^2 \qquad (7) \\ &= \sigma^2(1 + \theta_1^2 + \theta_2^2 + ...) \ \text{ (since } Cov(\varepsilon_t, \varepsilon_{t-k}) = 0 \text{ for all } k) \\ &= \sigma^2 \sum_{j=0}^{\infty} \theta_j^2 \ \text{ where } \theta_o = 1 \qquad (8) \end{aligned}$$

The covariance

$$\begin{aligned}
\rho_k &= E(x_t - \mu)(x_{t-k} - \mu) \\
&= E(\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ...)(\varepsilon_{t-k} + \theta_1\varepsilon_{t-1-k} + \theta_2\varepsilon_{t-2-k} + ...) \quad (9) \\
&= E(\theta_k\varepsilon_{t-k}\varepsilon_{t-k}) + E(\theta_{k+1}\theta_1\varepsilon_{t-k-1}\varepsilon_{t-k-1}) + ... \\
&= \sigma^2 \sum_{j=0}^{\infty} \theta_j\theta_{j+k}, \quad \text{where } \theta_1 = 0 \qquad (10)
\end{aligned}$$

Moving Average Processes

A moving average process of order q is a special case of equation (5) where the number of lags are truncated at q. For $y_t = x_t - \mu$, is written as

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ... + \theta_q\varepsilon_{t-q}, \ t = 1, ..., T$$

and denoted by $y_t \tilde{}\text{MA}(q)$

A finite moving average is always stationary since equations (6), (7), and (8) will automatically satisfy the weak stationary conditions for a finite sum.

Example MA(1)

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} \qquad (11)$$

Then

(i) $E(y_t) = 0$

(ii) $E(y_t)^2 = E(\varepsilon_t + \theta_1\varepsilon_{t-1})^2 = \sigma^2(1 + \theta_1^2)$

(iii)

$$\begin{aligned}
E(y_t y_{t-k}) &= E(\varepsilon_t + \theta_1\varepsilon_{t-1})(\varepsilon_{t-k} + \theta_1\varepsilon_{t-k-1}) \\
&\quad \begin{cases} \sigma^2\theta_1 & \text{for } k = 1 \\ 0 & \text{for } k > 1 \end{cases}
\end{aligned}$$

(iv)

$$\rho_k = \begin{cases} \theta_1/(1 + \theta_1^2) & \text{for } k = 1 \\ 0 & \text{for } k > 1 \end{cases}$$

Example MA(q)

(i) $E(y_t) = 0$

(ii) $E(y_t)^2 = \sigma^2(1 + \theta_1^2 + \theta_2^2 + ... + \theta_q^2)$

(*iii*)

$$E(y_t y_{t-k}) = \begin{cases} \sum_{j=0}^{q} \sigma^2 \theta_j \theta_{j+k} & \text{for } k = 1, 2, ..., q \\ 0 & \text{for } k > q \end{cases}$$

(*iv*)

$$\rho_k = \begin{cases} \sum_{j=0}^{q} \sigma^2 \theta_j \theta_{j+k} / \sum_{j=0}^{q} \theta_j^2 & \text{for } k = 1, 2, ..., q \\ 0 & \text{for } k > q \end{cases}$$

Autoregressive Model

An autoregressive process of order p is written as

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t, \quad t = 1, ..., T \tag{12}$$

This will be denoted $y_t \tilde{}$AR(p)

Example AR(1)

$$y_t = \phi_1 y_{t-1} + \varepsilon_t \qquad t = 1, ..., T \tag{13}$$

Notice that if this relationship is valid for time t, it should also be valid for time $t-1$, that is

$$y_{t-1} = \phi_1 y_{t-2} + \varepsilon_{t-1} \tag{14}$$

Substituting equation (12) into equation (11) we get the following expression.

$$\begin{aligned} y_t &= \phi_1(\phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \phi_1^2 y_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \end{aligned}$$

and repeating this procedure $j-1$ times we get

$$y_t = \phi_1^j y_{t-j} + \phi_1^{j-1} \varepsilon_{t-(j-1)} + \phi_1^{j-2} \varepsilon_{t-(j-2)} + ... + \phi_1 \varepsilon_{t-1} + \varepsilon_t \tag{15}$$

Now if $|\phi| < 1$ the deterministic component of y_t is negligible if j is large enough. Under this condition equation (13) might be written as

$$y_t = \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j} \tag{16}$$

In other words, whenever $|\phi_1| < 1$, an autoregressive process of order 1 may be written as an infinite moving average process in which the coefficient of ε_{t-j} is ϕ_1^j.

The first point to establish about an autoregressive process is the conditions under which it is stationary. Clearly, for the AR(1) process the condition for stationarity is $|\phi_1| < 1$ since whenever this condition holds the weak stationary conditions are automatically satisfied:

Proof:

(i) The mean exists and does not depend on time.

$E(y_t) = E(\sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}) = 0$

Notice that this is only true when $|\phi_1| < 1$.

Using equation (13), we can easily verify that when $|\phi_1| \geq 1$, then $E(y_t) = \phi_1^j y_{t-j}$

Therefore, whenever $|\phi_1| \geq 1$, the expected value of y_t depends on t, and then violates the stationarity condition.

(ii) The variance exists and does not depend on time.

$$\begin{aligned} V(y_t) &= V(\sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}) \\ &= E(\sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j})^2 && \text{(since } E(y_t) = 0\text{)} \\ &= E(\sum_{j=0}^{\infty} \phi_1^{2j} \varepsilon_{t-j}^2) && \text{(since } \varepsilon \text{ is a WN process)} \\ &= \sum_{j=0}^{\infty} \phi_1^{2j} E(\varepsilon_{t-j}^2) = \sigma^2 \sum_{j=0}^{\infty} \phi_1^{2j} = \frac{\sigma^2}{(1-\phi_1^2)} && \text{(since } |\phi_1| < 1\text{)} \end{aligned}$$

or

$$\gamma_0 = \frac{\sigma^2}{(1-\phi_1^2)} \tag{17}$$

(iii) The autocovariances exist and do not depend on time.

To calculate the autocovariance of an autoregressive process is slightly more complicated than that of a moving average. We proceed in the following way;

$$E(y_t y_{t-k}) = \phi_1 E(y_{t-1} y_{t-k}) + E(\varepsilon_t y_{t-k}),$$

or

$$\gamma_k = \phi_1 \gamma_{k-1} + E(\varepsilon_t y_{t-k})$$

Now notice that

$$E(\varepsilon_t y_{t-k}) = E[\varepsilon_t(\phi_1^{j-1} \varepsilon_{t-k-(j-1)} + \phi_1^{j-2} \varepsilon_{t-k-(j-2)} + ... + \phi_1 \varepsilon_{t-k-1)} + \varepsilon_{t-k})]$$

Given that the error terms are WN processes, this expression is equal to zero for $k > 0$, and we can write the autocovariance function as

$$\gamma_k = \phi_1 \gamma_{k-1} \tag{18}$$

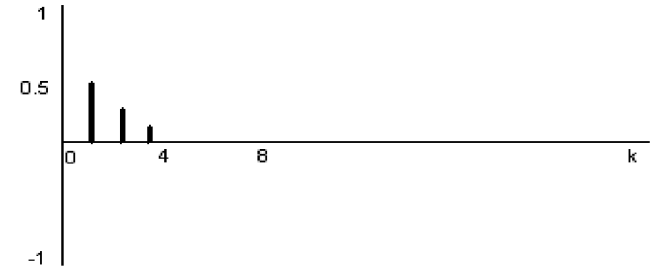
From equation (16) we can easily derive the autocorrelation function that is

$$\rho_k = \phi_1 \rho_{k-1} \tag{19}$$

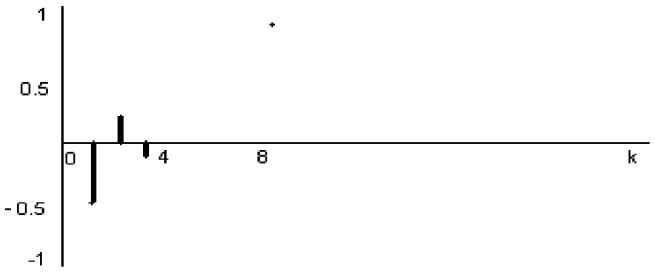
Therefore whenever the process is stationary the autocorrelation function declines exponentially. Using equation (17) it can easily be seen that $\rho_k = \phi_1^k \rho_0$.

Examples

i)$\phi_1 = .5$



ii)$\phi_1 = -.5$



The Lag Operator

The lag operator, L, is defined by the transformation

$$Ly_t = y_{t-1} \tag{20}$$

Notice that the lag operator may also be applied to y_{t-1} yielding

$$Ly_{t-1} = y_{t-2} \tag{21}$$

Now substituting (18) into (19) we get $Ly_{t-1} = L(Ly_t) = L^2 y_t = y_{t-2}$ and so in general

$$L^k y_t = y_{t-k} \qquad \text{for } k \geq 0 \tag{22}$$

The lag operator can be manipulated in a similar way to any algebraic quantity.

Example

Let us reproduce for convenience equation (14), an infinite moving average process in which the coefficient of ε_{t-j} is ϕ_1^j , that is,

$y_t = \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}$ where we assume $|\phi_1| < 1$, then using the lag operator this expression may be written as

$$y_t = \sum_{j=0}^{\infty} (\phi_1 L)^j \varepsilon_t = \varepsilon_t / (1 - \phi_1 L)$$

Notice that L is regarded as having the property that $|L| \leq 1$, and then $|\phi_1 L| < 1$, which is a necessary condition for the convergence of the series.

This can be rearranged in the following way

$$(1 - \phi_1 L) y_t = \varepsilon_t$$

or

$$y_t = \phi_1 y_{t-1} + \varepsilon_t$$

The Difference operator

The first difference operator, Δ, is defined as $\Delta = 1 - L$.

For example

$$\Delta y_t = (1 - L) y_t = y_t - y_{t-1}.$$

and

$$\Delta^2 y_t = (1 - L)^2 y_t = (1 - 2L + L^2) y_t = y_t - 2y_t + y_{t-2}$$

Autoregressive processes using Lag operators

An AR(p) process may be written as,

$$(1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) y_t = \varepsilon_t, \qquad t = 1, ..., T$$

or

$$\phi(L) y_t = \varepsilon_t, \qquad t = 1, ..., T$$

where $\phi(L) = (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p)$.

The stationarity condition for an autoregressive process may be expressed in terms of the roots of the polynomial of order p in L.

This may be easily understood for a first order autoregressive process. We have shown that an $AR(1)$ process may be written as,

$$(1 - \phi_1 L) y_t = \varepsilon_t, \qquad t = 1, ..., T$$

then we consider the roots (one in this case) of the polynomial in L, $(1 - \phi_1 L) = 0$, that is $L = 1/\phi_1$, which is greater than 1 (in absolute value) whenever $|\phi_1| < 1$.

In general an autoregressive process of order p is said to be stationary when all the roots of the polynomial $(1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p)$ lie outside the "unit circle". (there are all greater than one in absolute value).

Moving average processes using Lag operators

$$y_t = (1 + \theta_1 L + \theta_2 L^2 + ... + \phi_q L^q) \varepsilon_t \qquad t = 1, ..., T$$

or

$$y_t = \theta(L) \varepsilon_t$$

where $\theta(L) = (1 + \theta_1 L + \theta_2 L^2 + ... + \phi_q L^q)$.

Sometimes we want to express a moving average as an autoregressive process. For this to be possible we need to impose conditions on the parameters similar to the ones we impose for stationarity. If these conditions hold the moving average process is said to be *invertible*.

Invertibility

ARMA(q) process is said to be invertible if all the roots of the polynomial $(1 + \theta_1 L + \theta_2 L^2 + ... + \phi_q L^q)$ lie outside the unit circle.

Autoregressive Moving Average processes - ARMA processes

An autoregressive moving average process of order (p, q), denoted as ARMA(p, q) is written as

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} \qquad t = 1, ..., T$$

or

$$\phi(L) y_t = \theta(L) \varepsilon_t$$

with $\phi(L)$ and $\theta(L)$ defined as before.

Notice that the AR(p) and the MA(q) are special cases of the ARMA(p, q) process. The stationarity of an ARMA process depends solely on its autoregressive part and the invertibility only on its moving average part. Therefore an ARMA process is stationary if the roots of $\phi(L)$ are outside the unit circle and it is invertible whenever the roots of $\theta(L)$ are outside the unit circle. If both conditions hold an ARMA process can be written either as an infinite autoregressive process or as a infinite moving average process.

Example ARMA(1,1)

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Autocovariance function of an ARMA(1,1).

$$\begin{aligned} \gamma_k &= E(y_t y_{t-k}) = \phi_1 E(y_{t-1} y_{t-k}) + E(\varepsilon_t y_{t-k}) + \theta_1 E(\varepsilon_{t-1} y_{t-k}) \\ &= \phi_1 \gamma_{k-1} + E(\varepsilon_t y_{t-k}) + \theta_1 E(\varepsilon_{t-1} y_{t-k}) \end{aligned}$$

When $k = 0$

$$\gamma_0 = \phi_1 \gamma_1 + E(\varepsilon_t(\phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1})) + \theta_1 E(\varepsilon_{t-1}(\phi_1 y_{t-1} + \varepsilon_t + \phi_1 \varepsilon_{t-1})$$

where

- $E(\varepsilon_t(\phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1})) = \sigma^2$
- $E(\varepsilon_{t-1}(\phi_1(\phi_1 y_{t-2} + \varepsilon_{t-1} + \theta_1 \varepsilon_{t-2}) + \varepsilon_t + \theta_1 \varepsilon_{t-1})) = (\phi_1 + \theta_1)\sigma^2$

then

$$\gamma_0 = \phi_1 \gamma_1 + \sigma^2 + \theta_1(\phi_1 + \theta_1)\sigma^2$$

When $k = 1$

$$\gamma_1 = \phi_1 \gamma_0 + \theta_1 E(\varepsilon_{t-1}(\phi_1 y_{t-2} + \varepsilon_{t-1} + \theta_1 \varepsilon_{t-2}))$$

then

$$\gamma_1 = \phi_1 \gamma_0 + \theta_1 \sigma^2$$

for $k \geq 2$

$$\gamma_k = \phi_1 \gamma_{k-1}$$

The autocovariance function is therefore

(i) $\gamma_0 = \phi_1 \gamma_1 + \sigma^2 + \theta_1(\phi_1 + \theta_1)\sigma^2$

(ii) $\gamma_1 = \phi_1 \gamma_0 + \theta_1 \sigma^2$

(iii) $\gamma_k = \phi_1 \gamma_{k-1}$

Equations (i) and (ii) are a system of two equations with two unknowns γ_0 and γ_1 .

$$\gamma_0 = \frac{1 + \theta_1^2 + 2\theta_1\phi_1}{1 - \phi_1^2}\sigma^2$$

$$\gamma_1 = \frac{(1 + \theta_1\phi_1)(\phi_1 + \theta_1)}{1 - \phi_1^2}\sigma^2$$

Partial Autocorrelations

Autocorrelation functions are very useful to identify the existence and the order of a moving average processes. We have also shown that the autocorrelation function of an autoregressive process declines exponentially, but it is difficult to guess the **order** of the autoregressive process from the plot of the autocorrelation function. In other words we know that the autocorrelation function of an autoregressive process declines exponentially but this plot does not enables us to distinguish between an AR(p) and a AR($p+1$) process.

To help with this problem of discrimination we define the **partial autocorrelation function - PACF**. In general, the correlation between two random variables is due to both being correlated with a third variable, e.g the correlation between y_t and y_{t-2} , for an AR(1) process has to come through the correlation between y_t and y_{t-1} on the one hand and y_{t-1} and y_{t-2} on the other hand.

So the k^{th} partial autocorrelation, $\phi_k = \phi_{kk}$, function measures the correlation not accounted for by an AR($k-1$) process.

For an autoregressive process of order p the Yule-Walker equations are given by the following recursion formulae;

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \phi_3\rho_{k-3} + ... + \phi_p\rho_{k-p} \qquad \text{for } k = 1, ...p.$$

Then we just need to set $k = p$ or $\phi_p = \phi_{kk}$ and solve the following system of equations.

$$\rho_k = \phi_{11}\rho_{k-1} + \phi_{22}\rho_{k-2} + \phi_{33}\rho_{k-3} + ... + \phi_{kk}$$

Then I give values to k that range from 1 to k and generate a system of k equations in k unknowns.

$$\begin{aligned}
\rho_1 &= \phi_{11}\rho_0 + \phi_{22}\rho_1 + \phi_{33}\rho_2 + ... + \phi_{kk}\rho_{k-1} && \text{for } k = 1 \\
\rho_2 &= \phi_{11}\rho_1 + \phi_{22}\rho_0 + \phi_{33}\rho_1 + ... + \phi_{kk}\rho_{k-2} && \text{for } k = 2 \\
\rho_k &= \phi_{11}\rho_{k-1} + \phi_{22}\rho_{k-2} + \phi_{33}\rho_{k-3} + ... + \phi_{kk}\rho_0 && \text{for } k = k
\end{aligned}$$

And solve the system for ϕ_{kk} using kramer's rule.

In practice, however we are ignorant of the true values of ρ_i as well as k (the order of the autoregressive process), which is of course, the whole problem. As we will see later, the empirical methodology will consist in trying to find which ϕ_{kk} is not significantly different from zero.

If the process generating the data is of pure moving average form, what pattern would we expect to find for the partial autocorrelation function? Since an MA process may be written as an AR process of infinite order, we should expect a moving average process decays exponentially.

Example AR(1)

$\rho_k = \phi_1\rho_{k-1}$, then $\rho_k = \phi_{11}\rho_{k-1}$ since $p = k = 1$,
$\rho_1 = \phi_{11}\rho_0$ for $k = 1$.

Then

$$\begin{aligned}
\rho_1 &= \phi_1 = \phi_{11} && \text{for } k = 1. \\
\phi_{ii} &= 0 && \text{for } k > 1.
\end{aligned}$$

Example AR(2)

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2}$$

then

$$\rho_k = \phi_{11}\rho_{k-1} + \phi_{22}\rho_{k-2} \text{ since } p = k = 2,$$

Giving values to k we construct the following system of equations

$$\begin{aligned}
\rho_1 &= \phi_{11} + \phi_{22}\rho_1 && for\ k = 1, \\
\rho_2 &= \phi_{11}\rho_1 + \phi_{22} && for\ k = 2,
\end{aligned}$$

then

$$\phi_{22} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}.$$

$\phi_{ii} = 0$ for $k > 2$

Summary of Identification rules using ACF and PACf

For an AR(p) Process

(i) the ACF declines exponentially

(ii) the PACF is zero for lags greater than p.

For a MA(q) Process

(i) the ACF is zero for lags greater than q.

(ii) the PACF declines exponentially

Therefore using *sample* information, we might calculate sample ACF and PACF to try to identify the right model. These methodology advocated by Box and Jenkins usually consist of four steps.

(1) Transform the data, if necessary, so that the assumption of covariance stationarity is a reasonable one.

(2) Make an initial guess of small values of p and q for an ARMA(p, q) model that might describe the transformed series.

(3) Estimate the parameters in $\phi(L)$ and $\theta(L)$

(4) Perform diagnostic analysis to confirm that the model is indeed consistent with the observed features of the data.

We have up to now assumed (1) holds and described point (2). Now we are going to explain both the empirical properties of the sample analogs of the above defined parameters and how to estimate these models.

Properties of the correlogram, the PACF and other Sample Statistics.

The correlogram is the basic tool of analysis in the time domain. An inspection of the correlogram may lead to the conclusion that the series is random, or that exhibits a pattern of serial correlation that which perhaps can be modeled by a particular stochastic process. In order to decide which model is best representing the data, it is necessary to know something about the sampling properties of the correlogram and related statistics such as the mean and autocovariances.

Sample analogs for the ACF and PACF function

We have described the autocorrelation and partial autocorrelation function in terms of the population autocorrelations. These values can be estimated from a single series. In general this involves to calculate the following sample moments

The sample mean

The sample mean, $\hat{\mu}$, is an unbiased estimator of the mean of a stationary process, μ. It is calculated as

$$\hat{\mu} = T^{-1} \sum_{t=1}^{T} y_t$$

It can easily be shown that $\hat{\mu}$ is unbiased. It can also be shown that, although it is algebraically demanding, that the sample mean is also a consistent estimator.

The Sample Variance

$$\hat{\gamma}_0 = T^{-1} \sum_{t=1}^{T} (y_t - \hat{\mu})^2$$

The sample Autocovariances.

$$\hat{\gamma}_k = T^{-1} \sum_{t=k+1}^{T} (y_t - \hat{\mu})(y_{t-k} - \hat{\mu})$$

The sample Autocorrelations

The sample autocorrelation, $\hat{\rho}_k$, is defined as the ratio of $\hat{\gamma}_k$ and $\hat{\gamma}_0$.

$$\hat{\rho}_k = \hat{\gamma}_k / \hat{\gamma}_0$$

It can be shown that the asymptotic variance of $\hat{\rho}_k$, Avar$(\hat{\rho}_k)$, is approximately $(1/T)$, where T is the sample size. Using this approximation, the standard deviation is clearly $\sqrt{(1/T)}$.

Testing for the significance of ρ_k

In order to identify in practice, using autocorrelation functions, which particular type of model is the one that best represents the data, we should test whether the different parameters ρ_k are different from zero. Under the null hypothesis, i.e. $\rho_k = 0$, $\hat{\rho}_k$ is distributed asymptotically (valid for large samples) Normal with mean zero and variance $(1/T)$.

Proceeding on this basis, a test may be carried out on the sample autocorrelation at a particular lag, τ, by treating $\sqrt{T}\hat{\rho}_k$ as a standarised normal variable. At a five percent level of significance, the null hypothesis is rejected if the absolute value of $\sqrt{T}\hat{\rho}_k$ is greater than 1.96.

Testing for the significance of ϕ_{kk}

We proceed in a similar way than the one we describe to identify the particular model using autocorrelation functions. We test whether the different parameters ϕ_{kk} are different from zero.

Under the null hypothesis, i.e. $\phi_{kk} = 0$, is distributed approximately asymptotically Normal with mean zero and variance $(1/T)$.

Unfortunately these identifying tools won't tell us neither whether the preferred model is mispecified, nor what to do when two different models, say ARMA(1,2) and ARMA(2,1) seem to be equally valid. Therefore we will need to estimate these models.

Autoregressive models may be estimated simply by OLS but this procedure is not useful whenever the model has Moving Average terms. To estimate these models we need to use another procedure.

Maximum Likelihood Estimation

The principle on which estimation of ARMA models will be based is that of maximum likelihood.

We will present this principle for the simplest case which entails to find estimators of the mean and the variance of a random variable, say X, which is known to be normally distributed. The vector of population parameters is (μ, σ^2).

The principle may be expressed as follows. Given a sample $(x_1, x_2, ..., x_n)$, which are the values of the population parameters that have most likely generated that sample.

We then define the likelihood function as a function of the parameters given the sample, that is,

$$\mathcal{L}(\mu, \sigma^2 | x_1, x_2, ..., x_n) = f(x_1|\mu, \sigma^2) f(x_2|\mu, \sigma^2) ..... f(x_n|\mu, \sigma^2)$$

In writing the right hand side as the product of the density functions we have made use of two assumptions; i) the random variables are identically distributed, ii) there are independent.

We can rewrite the likelihood function as

$$\mathcal{L}(\mu, \sigma^2 | x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i | \mu, \sigma^2)$$

where Π is the multiplication operator and

$$f(x_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

Notice that

$$\prod_{i=1}^{n} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}}$$

The Maximum likelihood estimators, $\hat{\mu}$ and $\hat{\sigma}^2$, are designed to maximize the likelihood that the sample comes from a normal distribution with parameters μ and σ^2. To find them optimally we just differentiate the likelihood function with respect to μ and σ^2.

Notice that if we make a monotonic transformation of the likelihood function, the optimal values are not affected by the transformation. Sometimes is algebraically easier to maximize the logarithm of the maximum likelihood, that is

$$log(\mathcal{L}(\mu, \sigma^2 | x_1, x_2, ..., x_n)) = -\frac{n}{2} log(2\pi) - n log(\sigma) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}$$

Optimization

$$\frac{\partial log(\mathcal{L})}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^2} = 0$$

$$\frac{\partial log(\mathcal{L})}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^3} = 0$$

This system gives as solutions

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Conditional Maximum Likelihood Estimates

Usually when we estimate ARMA(p, q) models we evaluate the conditional maximum likelihood. What is meant by conditional is that we assume that the first $\max(p, q)$ observations are known. In practice we maximize the likelihood usually by numerical procedures.

For example for an AR(1) process

$$y_t = \phi_1 y_{t-1} + \varepsilon_t \qquad\qquad t = 1, 2, T$$

We then assume the log of joint distribution of $y_T, y_{T-1}, ..., y_2$ conditional on the value of y_1 .

$$f(y_T, y_{T-1}, ..., y_2 | y_1, \phi_1, \sigma^2) = \prod_{i=2}^{T} f(y_i | y_{i-1}, \phi_1, \sigma^2)$$

the objective then being to maximize

$$= -(T-1)log(2\pi)^{.5} - (T-1)\log\sigma - \frac{\sum_{t=2}^{T}(y_t - \phi_1 y_{t-1})^2}{2\sigma}.$$

We maximize this function by numerical procedures and obtain $\hat{\phi}_1$, $\hat{\sigma}^2$.

The "Portmanteau" statistic.

The final step in the Box - Jenkins methodology is to perform diagnostic analysis to confirm that the model is indeed consistent with the observed features of the data. If the model that we identified is the right one, the residuals of the estimated are supposes to be white noise. The most common test for whiteness of the residuals is the "Box and Pierce" test which makes use of the Q statistic.

The "Portmanteau" statistic, Q, defined as

$$Q^*(k) = T\sum_{i=1}^{k} \hat{\rho}_i^2$$

it can be shown to be asymptotically distributed, under the null hypothesis that y_t is a white noise, chi square with k degrees of freedom.

If the test is applied to the residuals of an estimated ARMA(p, q) model, say $\hat{\varepsilon}$, then $Q^*(k)$ is distributed $\chi^2(k - p - q)$.

This statistic has bad small sample properties. A better approximation is obtained by modifying the statistic in the following way.

$$Q(k) = T(T+2)\sum_{i=1}^{k}(T-i)^{-1}\hat{\rho}_i^2$$

This statistic is the one reported in the econometric package EVIEWS

.

The use of model Selection Criteria.

The model selection criteria is a set of rules that will help us to discriminate between alternative "successful" models. That is, it might be that we end with two alternative models that "pass" all the relevant test and I need to somehow to decide between them. The most used criteria are

The Akaike Criteria (AIC)

$$AIC(p,q) = \log \hat{\sigma}^2 + 2(p+q)T^{-1}$$

The Schwarz Criteria (BIC)

$$BIC(p,q) = \log \hat{\sigma}^2 + (p+q)T^{-1}\log(T)$$

These criteria are used as follows; whenever we have two different models that seem to be equally good we choose the model which has smallest AIC or BIC.

Forecasting with Time-Series Models

When we introduced the concept of stochastic process as models for time series at the beginning of the course, it was with the ultimate objective of using the models to infer from the past history of a series its likely course in the future. More precisely we want to derive from a model the conditional distribution of future observations given the past observations that it implies. This final step in the model building process is what we refer loosely as *forecasting*. It should be noted that in practice the model in hand is never the hypothetical "true" process generating the data we have observed. Rather, it is an approximation to the generating process and is subject to errors in both identification and estimation. Thus, although we shall discuss forecasting as if we knew the generating process, it is clear that our success in practice will depend in part on the adequacy of our empirical model and therefore on success in the preceding stages of identification and estimation.

Minimum Mean-square-error Forecasts

The main motivation for beginning the discussion about forecasting with the conditional expectation is that in many operational contexts it is desirable to be able to quote a point forecast, a single number, and the conditional expectation has the desirable property of being the *minimum mean square error forecast*. That is, if the model is correct, there is no other extrapolative forecast which will produce errors whose squares have smaller expected value.

Although we have not discussed how conditional expectations are computed, this general result is easily demonstrated as follows.

Given the availability of a set of observations up to, and including y_T , the *optimal predictor* l steps ahead is *the expected value of* y_{t+l} *conditional on the information at time* $t = T$. This may be written as

$$\widehat{y}_{t+l|T} = E^*(y_{t+l}|I_T)$$

The predictor is optimal in the sense that has minimum mean square error. This is easily seen by observing that for any predictor, $E(y_{t+l}|I_T)$, constructed on the basis of the information available at time T, the forecasting error can be split into parts:

$$y_{t+l} - \widehat{y}_{t+l|T} = [y_{t+l} - E(y_{t+l}|I_T)] + [E(y_{t+l}|I_T) - \widehat{y}_{t+l|T}]$$

Since the second term on the right hand side is fixed at time T, it follows that, on squaring the whole expression and taking expectations at time T, the cross-product term disappears leaving.

$$MSE(\widehat{y}_{t+l|T}) = Var(y_{t+l|T}) + [\widehat{y}_{t+l|T} - E(y_{t+l}|I_T)]^2$$

In the first term on the right hand side, the conditional variance of y_{t+l} , does not depend on $\widehat{y}_{t+l|T}$. Hence the minimum mean square estimate (MMSE) of y_{t+l} is given by the conditional mean and it is unique.

Computation of Conditional Expectation Forecasts

One-Step-Ahead Forecasts

We now consider the question of how to construct an MMSE of a future observation from an ARMA process, given observations up to and including time T. The ARMA process is assumed to be stationary and invertible, with known parameters and independent disturbances with mean zero and constant variance σ^2.

The equation of an ARMA(p, q) model at time T+1 is

$$\begin{aligned} y_{T+1} = \quad & \phi_1 y_T + \phi_2 y_{T-1} + ... + \phi_p y_{T-p+1} + \\ & \varepsilon_{T+1} + \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1} + ... + \theta_q \varepsilon_{T-q+1} \end{aligned}$$

then

$$\begin{aligned} \widehat{y}_{t+1|T} = \quad & \phi_1 y_T + \phi_2 y_{T-1} + ... + \phi_p y_{T-p+1} + \\ & \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1} + ... + \theta_q \varepsilon_{T-q+1} \end{aligned}$$

Since all variables with time subscripts through period T have been realised (are no longer random) and $E(\varepsilon_{T+1}|I_T) = 0$.

For the numerical evaluation of $\widehat{y}_{t+l|T}$ from the above equation we need a value for the disturbances.

Optimal predictions for ARMA models

We now consider the question of how to construct an MMSE of a future observation from an ARMA process, given observations up to and including time T. The ARMA process is assumed to be stationary and invertible, with known parameters and independent disturbances with mean zero and constant variance σ^2.

The equation of an ARMA(p,q) model at time $T+l$ is

$$\begin{aligned}\widehat{y}_{T+l|T} = \quad & \phi_1\widehat{y}_{T+l-1|T} + \phi_2\widehat{y}_{T+l-2|T} + ... + \phi_p\widehat{y}_{T+l-p|T} + \\ & \varepsilon_{T+l|T} + \theta_1\hat{\varepsilon}_{T+l-1|T} + \theta_2\hat{\varepsilon}_{T+l-2|T} + ... + \theta_q\hat{\varepsilon}_{T+l-q}\end{aligned}$$

$$l = 1, 2, ...$$

Where

$\widehat{y}_{T+j|T} = y_{T+j}$ for $j \leq 0$ and $\hat{\varepsilon}_{T+j|T} = \begin{cases} 0 & \text{for } j > 0 \\ \varepsilon_{t+j} & \text{for } j \leq 0 \end{cases}$.

This expression provides a recursion for computing optimal predictions of the future observations.

Example 1 For the AR(1) process

$$y_{T+l} = \phi_1 y_{T+l-1} + \varepsilon_{T+l} \qquad \textit{at time } T+l$$

$$\widehat{y}_{T+l|T} = \phi_1\widehat{y}_{T+l-1|T} \qquad l = 1, 2, ...$$

The starting value is given by $\widehat{y}_{T|T} = y_T$, and so the previous equation may be solved to yield

$\widehat{y}_{T+l|T} = \phi_1^l y_T$

thus the predicted values decline exponentially towards zero, and the forecast function has exactly the same form as the autocovariance function.

Let us calculate **the forecast error** for this process

$$\begin{aligned} y_{T+l} - \widehat{y}_{T+l|T} &= \phi_1 y_{T+l-1} + \varepsilon_{T+l} - \phi_1^l y_T \\ &= \phi_1^l y_T + \varepsilon_{T+l} + \phi_1\varepsilon_{T+l-1} + \phi_1^2\varepsilon_{T+l-2} + ... + \phi_1^{l-1}\varepsilon_{T+l} - \phi_1^l y_T \end{aligned}$$

Then, the variance of the forecast error l periods ahead is given by

$$\begin{aligned} V(y_{T+l} - \widehat{y}_{T+l|T}) &= V(\varepsilon_{T+l} + \phi_1\varepsilon_{T+l-1} + \phi_1^2\varepsilon_{T+l-2} + ... + \phi_1^{l-1}\varepsilon_{T+l}) \\ &= (1 + \phi_1^2 + \phi_1^4 + ... + \phi_1^{2(l-1)})\sigma^2 \end{aligned}$$

Note that the variance of the forecast error increases (nonlinearly) as l becomes large.

Example 2

At time $T+1$, the equation for an MA(1) process is of the form

$$y_{T+1} = \varepsilon_{T+1} + \theta_1\varepsilon_T$$

Then in general

$$\widehat{y}_{T+l|T} = \hat{\varepsilon}_{T+l|T} + \theta_1 \hat{\varepsilon}_{T+l-1|T}$$

$$\begin{aligned} \widehat{y}_{T+l|T} &= \theta_1 \varepsilon_T \qquad for\ l = 1 \\ &= 0 \qquad for\ l > 1. \end{aligned}$$

The variance of the forecast error for a MA(1) is

$$\begin{aligned} V(y_{T+l} - \widehat{y}_{T+l|T}) &= \sigma^2 \qquad \text{for } l = 1 \\ &= (1 + \theta_1^2)\sigma^2 \qquad \text{for } l > 1 \end{aligned}$$

Thus the forecast error variance is the same for a forecast 2, 3, etc periods ahead, etc.

The ARMA(1,1) Process

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

$$\begin{aligned} \widehat{y}_{T+l|T} &= \phi_1 y_T + \theta_1 \varepsilon_T \qquad \text{for } l = 1 \\ &= \phi_1 \widehat{y}_{T+l-1|T} \qquad \text{for } l > 1 \\ &= \phi_1^l y_T + \phi_1^{l-1} \theta_1 \varepsilon_T \end{aligned}$$

(derive the MSE of the forecast as before).

Measuring the Accuracy of Forecasts

Various measures have been proposed for assessing the predictive accuracy of forecasting models. Most of these measures are designed to evaluate ex-post forecasts. The most well known are

The Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{l} \sum_{i=T+1}^{T+l} (\widehat{Y}_i - Y_i)^2}$$

where l is the number of periods being forecasted

The Mean Absolute Error.

$$MAE = \frac{1}{l} \sum_{i=T+1}^{T+l} |\widehat{Y}_i - Y_i|$$

Which indicator should be used depends of the purpose of the forecasting exercise. The RMSE will penalize big errors more than the MAE measure.

Consider the following two models, say 1 and 2. Assume model 1 forecasts accurately most of the time but performs very badly for an unusual observation. On the other hand assume that model 2 forecasting performance is poor most of the time but predicts the unusual observation with small error. Comparing the forecasting performances of these models whenever we use the RMSE indicator we would probably favour model 2, and favour model 1 when the MAE criteria is used. In this extreme experiment the preferred model depends very much on the preferences of the user, that is whether she prefers to forecast most of the time poorly but get right the unusual observation (e.g. a devaluation of the currency) or have most of the time a good forecast even if forecasts completely bad the unusual observation (buy pounds on tuesday before black Wednesday).

Appendix 1

White's Theorem

Theorem 1 *If* $\{Y_t\}_{t=1}^{\infty}$ *is a martingale difference sequence with* $\overline{Y}_T = \frac{1}{T}\sum Y_t$ *and*

- $E(Y_t^2) = \sigma_t^2$ *with* $\sigma_t^2 \xrightarrow{P} \sigma^2$
- *The moments* $E|Y_t|^r$ *exist for* $r \geq 2$
- $\frac{1}{T}\sum Y_t^2 \xrightarrow{P} \sigma^2$

then it can be shown that $\sqrt{T}\overline{Y}_T \tilde{} N(0, \sigma^2)$.

Consider now the following infinite moving average representation for Y_t,

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

with $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, and define the randomvariable $X_t = \varepsilon_t Y_{t-k}$ for $k > 0$. Then, X_t is a martingale difference (if ε_t is *iid*, $\varepsilon_t \varepsilon_{t-1}$ is a martingale difference) with variance $E(X_t^2) = \sigma^2 E(Y_t^2)$ and fourth moment $E(\varepsilon_t^4)E(Y_t^4) < \infty$.

Now if we can prove that $\frac{1}{T}\sum X_t^2 \xrightarrow{P} E(X_t^2)$ we would be under the conditions of the White theorem and can use that $\sqrt{T}\overline{X}_T = \frac{1}{\sqrt{T}}\sum X_t \tilde{} N(0, E(X_t^2))$. or alternatively

$$\frac{1}{\sqrt{T}}\sum \varepsilon_t Y_{t-k} \tilde{} N(0, \sigma^2 E(Y_t^2))$$

Proposition 2 $\frac{1}{T}\sum X_t^2 \xrightarrow{P} E(X_t^2)$:

Proof. To prove proposition first note that $\frac{1}{T}\sum X_t^2 = \frac{1}{T}\sum \varepsilon_t^2 Y_{t-k}^2 = \frac{1}{T}\sum(\varepsilon_t^2 - \sigma^2)Y_{t-k}^2 + \frac{1}{T}\sum \sigma^2 Y_{t-k}^2 \xrightarrow{P} \sigma^2 E(Y_t^2)$.This results arise since

(**i**) $(\varepsilon_t^2 - \sigma^2)Y_{t-k}^2$ is a martingale difference with finite second moments and therefore $\frac{1}{T}\sum(\varepsilon_t^2 - \sigma^2)Y_{t-k}^2 \xrightarrow{P} 0$,

(**ii**) $\frac{1}{T}\sum \sigma^2 Y_{t-k}^2 \xrightarrow{P} \sigma^2 E(Y_t^2)$.

Then, it follows that $\frac{1}{T}\sum X_t^2 \xrightarrow{P} \sigma^2 E(Y_t^2) = E(X_t^2)$. ■

Asymptotics of an AR(p) process.

Consider an autoregressive process

$$x_t = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \varepsilon_t$$

We may write the standard autoregressive model in regression notation

$$y_t = z_t\beta + u_t$$

where $y_t = x_t, \quad z_t = \{1, x_{t-1}, x_{t-2}, ...x_{t-p}\}\prime$, etc.

Here we cannot assume u_t is independent of z_{t+1} , although is independent of z_t . Without this we cannot apply any of the small sample results and have to rely on asymptotic results.

Consider the OLS estimator of β. Then we can write

$$\sqrt{T}(b_T - \beta) = ((1/T)\sum z_t z_t')^{-1}((1/\sqrt{T})\sum z_t u_t)$$

where

$$((1/T)\sum z_t z_t')^{-1} = \begin{bmatrix} 1 & T^{-1}\sum x_{t-1} & . & T^{-1}\sum x_{t-1} \\ T^{-1}\sum x_{t-1} & T^{-1}\sum x_{t-1}^2 & . & T^{-1}\sum x_{t-1}x_{t-p} \\ . & . & . & . \\ T^{-1}\sum x_{t-p} & T^{-1}\sum x_{t-p}x_{t-1} & . & T^{-1}\sum x_{t-p}^2 \end{bmatrix}^{-1}$$

The elements of the first row converge in probability to $\mu = E(x_t)$ and $T^{-1}\sum x_{t-i}x_{t-j}$ converges in probability to $E(x_{t-i}x_{t-j}) = \gamma_{i-j} + \mu^2$

Then $((1/T)\sum z_t z_t')^{-1}$ converges in probability to Q^{-1}, with the elements of Q defined as above.

For the second term $z_t u_t$ is a martingale difference with positive definite variance covariance given by $E(z_t u_t u_t z_t') = E(u_t^2)E(z_t z_t') = \sigma^2 Q$.

Then using standard arguments

$$((1/\sqrt{T})\sum z_t u_t) \quad \underrightarrow{L} \quad N(0, \sigma^2 Q)$$

(notice that $p\lim \frac{1}{T}\sum var(z_t u_t) = \sigma^2 Q$ since $z_t u_t$ sequence of random vectors with $E(z_t u_t) = 0$ (a martingale difference) and $(z_t u_t u_t z_t') = E(u_t^2)E(z_t z_t') = \sigma^2 Q$.)

Then it follows that

$$\sqrt{T}(b_T - \beta) \quad \underrightarrow{L} \quad N(0, \sigma^2 Q^{-1})$$

(since $((1/T)\sum z_t z_t')^{-1} \xrightarrow{P} Q^{-1}$ and $\sqrt{T}(b_T - \beta) = (1/T)\sum z_t z_t')^{-1}((1/\sqrt{T})\sum z_t u_t) \underrightarrow{L} \quad N(0, \sigma^2 Q^{-1} Q Q^{-1\prime})$ $N(0, \sigma^2 Q^{-1})$.)

For an AR(1).

$$y_t = \phi_1 y_{t-1} + \varepsilon_t$$

Then $Q = E(y_{t-1}^2) = \gamma_0 = \sigma^2/(1-\phi^2)$.

and

$\sqrt{T}(\hat{\phi}_T - \phi) \underrightarrow{L} N(0, \sigma^2(\sigma^2/(1-\phi^2))^{-1}) = N(0, (1-\phi^2))$

Appendix 2

Forecasts based on Linear Projections and Updating these Projections.

Consider

$$P(Y_{T+1}|X_t) = \alpha\prime X_t$$

Then, if

$$E[(Y_{t+1} - \alpha\prime X_t)X_t'] = 0,$$

$\alpha\prime X_t$ is called a linear projection of Y_{t+1} on X_t .

Properties of linear projections

(*i*) $E(Y_{t+1}X_t') = \alpha\prime E(X_t X_t')$

then

$\alpha\prime = E(Y_{t+1}X_t')(E(X_t X_t'))^{-1}$

(*ii*) The mean square error associated with a linear projection is given by $E(Y_{t+1} - \alpha\prime X_t)^2 = E(Y_{t+1})^2 - E(Y_{t+1}X_t)(E(X_t X_t'))^{-1}E(X_t Y_{t+1})$ (once we substitute and rearrange terms)

(*iii*) If X_t includes a constant, then projection of $aY_{t+1} + b$ on X_t

$$P(aY_{t+1} + b|X_t) = aP(Y_{t+1}|X_t) + b$$

Updating a linear Projection and Triangular Factorizations

a) Triangular Factorizations

Consider the following Matrix

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{bmatrix}$$

Assume Ω is symmetric.

Now multiply the first row by $\Omega_{21}\Omega_{11}^{-1}$ and subtracting the result from the second row it yields a zero in $(2,1)$, while multiplying the first row by $\Omega_{31}\Omega_{11}^{-1}$and subtracting the result from the third row it yields a zero in $(3,1)$

Then if we pre-multiply by

$$E_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\Omega_{21}\Omega_{11}^{-1} & 1 & 0 \\ -\Omega_{31}\Omega_{11}^{-1} & 0 & 1 \end{bmatrix}$$

$$E_1\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ 0 & \underbrace{\Omega_{22}\text{-}\Omega_{21}\Omega_{11}^{-1}\Omega_{12}}_{h_{22}} & \underbrace{\Omega_{23}\text{-}\Omega_{21}\Omega_{11}^{-1}\Omega_{13}}_{h_{23}} \\ 0 & \underbrace{\Omega_{32}\text{-}\Omega_{31}\Omega_{11}^{-1}\Omega_{12}}_{h_{32}} & \underbrace{\Omega_{33}\text{-}\Omega_{31}\Omega_{11}^{-1}\Omega_{13}}_{h_{33}} \end{bmatrix}$$

Then

$$E_1\Omega E_1' = \begin{bmatrix} \Omega_{11} & 0 & 0 \\ 0 & \Omega_{22}\text{-}\Omega_{21}\Omega_{11}^{-1}\Omega_{12} & \Omega_{23}\text{-}\Omega_{21}\Omega_{11}^{-1}\Omega_{13} \\ 0 & \Omega_{32}\text{-}\Omega_{31}\Omega_{11}^{-1}\Omega_{12} & \Omega_{33}\text{-}\Omega_{31}\Omega_{11}^{-1}\Omega_{13} \end{bmatrix} = H$$

$$H = \begin{bmatrix} h_{11} & 0 & 0 \\ 0 & h_{22} & h_{23} \\ 0 & h_{32} & h_{33} \end{bmatrix}$$

Repeating the same line of reasoning let define

$$E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \text{-}h_{32}h_{22}^{-1} & 1 \end{bmatrix}$$

$$E_2H = \begin{bmatrix} h_{11} & 0 & 0 \\ 0 & h_{22} & h_{23} \\ 0 & 0 & h_{33}\text{-}h_{32}h_{22}^{-1}h_{23} \end{bmatrix}$$

and

$$E_2 H E_2' = \begin{bmatrix} h_{11} & 0 & 0 \\ 0 & h_{22} & 0 \\ 0 & 0 & h_{33}\text{-}h_{32}h_{22}^{-1}h_{23} \end{bmatrix} = D$$

Then Ω can always been written in the following way $\Omega = ADA\prime$ where $A = (E_2 E_1)^{-1} = E_1^{-1} E_2^{-1}$.

Where

$$E_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 \\ \Omega_{31}\Omega_{11}^{-1} & 0 & 1 \end{bmatrix},$$

$$E_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & h_{32}h_{22}^{-1} & 1 \end{bmatrix} \text{ and}$$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 \\ \Omega_{31}\Omega_{11}^{-1} & h_{32}h_{22}^{-1} & 1 \end{bmatrix}$$

Updating a Projection

Let $Y = \{Y_1, Y_2, ...Y_n\}'$be a vector of random variables whose second moment is

$\Omega = E(YY')$

Let $\Omega = ADA'$ be a triangular factorization of Ω and define W

$W = A^{-1}Y$

Then $E(WW') = D$, and the W are random variables which are uncorrelated.

Consider $n = 3$

Then

$$\begin{bmatrix} 1 & 0 & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 \\ \Omega_{31}\Omega_{11}^{-1} & h_{32}h_{22}^{-1} & 1 \end{bmatrix} \begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$$

The first equation states

$W_1 = Y_1$

The second equation $\Omega_{21}\Omega_{11}^{-1}W_1 + W_2 = Y_2$, and defining $\alpha = \Omega_{21}\Omega_{11}^{-1}$ and using the first equation we have $E(W_2 W_1) = 0$ (because of the orthogonalization) $= E((Y_2 - \alpha Y_1)Y_1) = 0$.

Then the triangular factorization can be used to infer the coefficient of a linear projection. In general row i of A has the interpretation of a linear projection of Y_i on Y_1 .

Then W_2 has the interpretation of the residual of a linear projection of Y_2 on Y_1 so its MSE is $E(W_2 W_2') = D_{22} = \Omega_{22}\text{-}\Omega_{21}\Omega_{11}^{-1}\Omega_{12}$

The third equation states that

$\Omega_{31}\Omega_{11}^{-1}W_1 + h_{32}h_{22}^{-1}W_2 + W_3 = Y_3$

or

$W_3 = Y_3 - \Omega_{31}\Omega_{11}^{-1}Y_1 - h_{32}h_{22}^{-1}(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1)$

Thus W_3 is the residual of subtracting some linear combination of Y_1 and Y_2 from Y_3 , and this residual is uncorrelated by construction with either W_1 or W_2 , $E(W_3W_1) = E(W_3W_2) = 0$.

Then

$E[(Y_3 - \Omega_{31}\Omega_{11}^{-1}Y_1 - h_{32}h_{22}^{-1}(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1))W_i]$ $\qquad i = 1$ or 2.

Then the linear projection is

$P(Y_3|Y_2, Y_1) = \Omega_{31}\Omega_{11}^{-1}Y_1 + h_{32}h_{22}^{-1}(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1)$

with $MSE = D_{33} = h_{33}$-$h_{32}h_{22}^{-1}h_{23}$

This last expression gives a convenient formula for updating a linear projection. Suppose we want to forecast Y_3 and have initial information about Y_1 .

Then

$P(Y_3|Y_1) = \Omega_{31}\Omega_{11}^{-1}Y_1$.

Let Y_2 represent some new information with which we want to update the forecast. If we where just asked the magnitude of Y_2 on the basis of Y_1 alone we get

$P(Y_2|Y_1) = \Omega_{21}\Omega_{11}^{-1}Y_1$.

On the other hand we know that

$P(Y_3|Y_1, Y_2) = \Omega_{31}\Omega_{11}^{-1}Y_1 + h_{32}h_{22}^{-1}(y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1)$

Then

$$P(Y_3|Y_1, Y_2) = P(Y_3|Y_1) + h_{32}h_{22}^{-1}(y_2 - P(Y_2|Y_1))$$

so we can thus optimally update the forecast $P(Y_3|Y_1)$ by adding to it a multiple $h_{32}h_{22}^{-1}$ of the unanticipated component of the new information.

Notice that $h_{22} = E(Y_2 - P(Y_2|Y_1))^2$ and $h_{32} = E(Y_2 - P(Y_2|Y_1))(Y_3 - P(Y_3|Y_1))$, then the projection formulae might be written as.

$$P(Y_3|Y_1, Y_2) = P(Y_3|Y_1)$$
$$+E(Y_2 - P(Y_2|Y_1))(Y_3 - P(Y_3|Y_1)).(E(Y_2 - P(Y_2|Y_1))^2)^{-1}(Y_2 - P(Y_3|Y_1)).$$

Vector Autoregressions

A vector autoregressive (VAR) is simply an autoregressive process for a vector of variables.

Let us define $W_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix}$, a matrix A_{2X2} and $\varepsilon_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$.

Then a **VAR(1)** may be written as

$$W_t = AW_{t-1} + \varepsilon_t$$

or

$$x_t = a_{11}x_{t-1} + a_{12}y_{t-1} + \varepsilon_{1t},$$
$$y_t = a_{21}x_{t-1} + a_{22}y_{t-1} + \varepsilon_{2t},$$

where

$$E(\varepsilon_t) = 0, \qquad E(\varepsilon_t\varepsilon_s') = \begin{cases} \Omega & t = s \ (\Omega = \Omega', \ c'\Omega c > 0, \ c \neq 0), \\ 0 & \text{otherwise.} \end{cases}$$

VAR(p)

$$W_t = A_1W_{t-1} + A_2W_{t-2} + ... + A_pW_{t-p} + \varepsilon_t$$

or

$$(I - A_1L - A_2L^2 - ... - A_pL^p)W_t = \varepsilon_t$$

The VAR is covariance stationary if all the values of L satisfying $|I - A_1L - A_2L^2 - ... - A_pL^p| = 0$ lie outside the unit circle[1].

The Autocovariance Matrix

For a covariance stationary n dimensional vector process we may define the *autocovariance function* for a VAR in a way similar to the univariate case

$$\Gamma_{k \ (\text{nxn})} = E(W_tW_{t-k}') \quad where \quad \Gamma_{k(ij)} = cov(W_{i,t}, W_{j,t-k})$$

Using the above two variables VAR we get the following

[1] For the VAR(1) example with 2 variables this is equivalent to say that the real part both roots in L are greater than 1.

$$\text{Det}\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} L \right] = 0$$

or

$$1 - (a_{11} + a_{22})L + (a_{11}a_{22} - a_{12}a_{21})L^2 = 0$$

$$\Gamma_{k\ (\text{nxn})} = E(W_t W'_{t-k}) = \begin{bmatrix} E(x_t x_{t-k}) & E(x_t y_{t-k}) \\ E(y_t x_{t-k}) & E(y_t y_{t-k}) \end{bmatrix}$$

Contrary to the univariate case $\Gamma_k \neq \Gamma_{-k}$, instead the correct relationship is $\Gamma'_k = \Gamma_{-k}$. This can be easily understood looking at the simple bivariate case where there is no reason why $E(x_t y_{t-1})$ should be equal to $E(x_{t-1} y_t)$.

Proof

i) Leading the vector k times we obtain

$$\Gamma_{k\ (\text{nxn})} = E(W_{t+k} W'_t)$$

ii) Taking transposes

$$\Gamma'_{k\ (\text{nxn})} = E(W_t W'_{t+k}) = \Gamma_{-k}$$

In terms of our two variables example

$$\Gamma_{k\ (\text{nxn})} = \begin{bmatrix} E(x_t x_{t-k}) & E(x_t y_{t-k}) \\ E(y_t x_{t-k}) & E(y_t y_{t-k}) \end{bmatrix};$$

$$\Gamma'_{k\ (\text{nxn})} = \begin{bmatrix} E(x_t x_{t-k}) & E(y_t x_{t-k}) \\ E(x_t y_{t-k}) & E(y_t y_{t-k}) \end{bmatrix} \overbrace{=}^{\textit{Leading } k \textit{ periods}} \begin{bmatrix} E(x_{t+k} x_t) & E(y_{t+k} x_t) \\ E(x_{t+k} y_t) & E(y_{t+k} y_t) \end{bmatrix};$$

$$\Gamma_{-k\ (\text{nxn})} = \begin{bmatrix} E(x_t x_{t+k}) & E(x_t y_{t+k}) \\ E(y_t x_{t+k}) & E(y_t y_{t+k}) \end{bmatrix}.$$

Autocovariance Function

We can now calculate the autocovariance function for a VAR more or less in the same way we did for the univariate process. Assume for simplicity that we have a first order VAR.

Then the autocovariance function can be derived as

$$\Gamma_{k\ (\text{nxn})} = E(W_t W'_{t-k}) = AE(W_{t-1} W'_{t-k}) + E(\varepsilon_t W'_{t-k})$$

then we obtain that for $k \geq 1$

$$\Gamma_{k\ (\text{nxn})} = A\Gamma_{k-1}.$$

When $k = 0$ we can simply notice that

$$E(W_t W'_t) = AE(W_{t-1} W'_{t-1})A' + E(\varepsilon_t \varepsilon'_t)$$

or

$$\Gamma_0 = A\Gamma_0 A' + \Omega$$

To obtain Γ_0 we use the fact that $vec(ABC) = (C' \otimes A)vec(B)$, then

$$vec(\Gamma_0) = vec(A\Gamma_0 A') + vec(\Omega) = (A \otimes A)vec(\Gamma_0) + vec(\Omega),$$

or

$$vec(\Gamma_0) = (I_{(np)^2} - (A \otimes A))^{-1} vec(\Omega).$$

The Companion Form

Notice that a VAR(p) may always be re-written as a VAR(1) by defining a vector H_t such that

$$H_t = FH_{t-1} + \nu_t$$

where

$$H_t = \begin{bmatrix} x_t \\ y_t \\ . \\ x_{t-i} \\ y_{t-i} \\ . \\ x_{t-(p-1)} \\ y_{t-(p-1)} \end{bmatrix}, \quad F = \begin{bmatrix} \mathrm{A}_1 & \mathrm{A}_2 & | & & \mathrm{A}_p \\ \mathrm{I}_{2\times 2} & 0 & | & & 0 \\ & \mathrm{I}_{2\times 2} & & & \\ 0 & 0 & \mathrm{I}_{2\times 2} & & 0 \\ & & | & \mathrm{I} & 0 \end{bmatrix} \quad and\, \nu_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ . \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Then, the VAR in the Companion form can be expressed in the following way

$$H_t = FH_{t-1} + \nu_t \qquad\qquad E(\nu_t \nu_t') = \begin{cases} Q & t = s \\ 0 & \text{otherwise} \end{cases}$$

where

$$Q_{(\mathrm{np\ x\ np})} = \begin{bmatrix} \Omega & 0 & 0 & ... & 0 \\ 0 & . & & & \\ . & . & & & \\ 0 & 0 & & ... & 0 \end{bmatrix}$$

Notice also that the following relationships holds:

$$E(H_t H_t') = FE(H_{t-1} H_{t-1}')F' + Q$$

or

$$\Sigma = F\Sigma F' + Q \qquad \text{where } \Sigma = E(H_t H_t').$$

where

$$\Sigma = \begin{bmatrix} \Gamma_0 & \Gamma_1 & ... & \Gamma_{p-1} \\ \Gamma_1' & \Gamma_0 & & \Gamma_{p-2} \\ \\ \Gamma_{p-1}' & \Gamma_{p-2}' & & \Gamma_0 \end{bmatrix}$$

and Γ_p is the autocovariance of the original process.

If the process is covariance stationary, then the unconditional variance can be calculated simply using vec operators, i.e.,

$$vec(\Sigma) = vec(F\Sigma F') + vec(Q) = (F \otimes F)vec(\Sigma) + vec(Q),$$

(Since $vec(ABC) = (C' \otimes A)vec(B)$)

Then the unconditional variance can be obtained as

$$vec(\Sigma) = (I_{(np)^2} - (F \otimes F))^{-1} vec(Q).$$

(NB $F \otimes F$ has no unit eigen values since the eigen values of $F \otimes F$ are of the form $\lambda_j \lambda_i$, and we knew that all $|\lambda_i| < 1$)

Notice as well that the j^{th} autocovariance function of H (denoted Σ_j) can be found by post-multiplying by H_{t-j}' and taking expectations.

$$E(H_t H_{t-j}') = FE(H_{t-1} H_{t-j}') + E(\nu_t H_{t-j}')$$

Thus

$$\Sigma_k = F\Sigma_{k-1} \quad \text{for } k = 1, 2, ...$$

or

$$\Sigma_k = F^k \Sigma.$$

The k^{th} autocovariance Γ_k of the original process W_t is given by the n first rows and n columns of $\Sigma_k = F\Sigma_{k-1}$:

$$\Gamma_k = A_1\Gamma_{k-1} + A_2\Gamma_{k-2} + ... + A_p\Gamma_{k-p} \qquad k = p, p+1, p+2, ...$$

Maximum likelihood Estimation.
The Conditional likelihood for a vector autoregression.

Let W_t denote an $(n_\times 1)$ vector containing the values that n variables take at time t. W_t is assumed to follow a p^{th} order gaussian VAR.

$$W_t = A_1 W_{t-1} + A_2 W_{t-2} + ... + A_p W_{t-p} + \varepsilon_t \qquad \varepsilon_t \tilde{} N(0, \Omega),$$

where both Ω and A_i are $n_\times n$ matrices of parameters.

Suppose that we observe these variables for $T + p$ time periods. The approach is to condition on the first p observations $(W_0, ...W_{-p+1})$ and to base the estimation on the last T observations $(W_T, ...W_1)$.

$$f(W_T, W_{T-1}, W_{T-2}....W_1 | W_0, ., W_{-p+1}; \Theta)$$

and maximize with respect to Θ, where Θ is a vector that contains the elements of $A_1, A_2, A_3, ...A_p$ and Ω.

Then

$$W_t | W_{t-1}, W_{t-2}, ...W_{-p+1} \tilde{} N((A_1 W_{t-1} + A_2 W_{t-2} + ... + A_p W_{t-p})_{n\times 1}, \Omega_{n\times n})$$

It will be convenient to stack the p lags in a vector x_t,.

$$x_t = \begin{bmatrix} \underbrace{W_{t-1}}_{nx1} \\ \underbrace{W_{t-2}}_{nx1} \\ . \\ \underbrace{W_{t-p}}_{nx1} \end{bmatrix} \quad np_\times 1$$

and let Π' denote the following $n_\times np$ matrix ; $\Pi' = [A_1, A_2, A_3, ...A_p]_{n\times np}$, then we can write the conditional mean as $\Pi' x_t$.

Using this notation, we can write the conditional distribution of W_t as $W_t | W_{t-1}, W_{t-2}, ...W_{-p+1} \tilde{} N(\Pi^{'} x_t, \Omega)$

$$\begin{aligned} & f(W_t | W_{t-1}, W_{t-2}....W_1, W_0, .., W_{-p+1}; \Theta) \\ = \quad & (2\pi)^{-n/2} |\Omega^{-1}|^{.5} \exp[(-1/2)(W_t - \Pi' x_t)' \Omega^{-1} (W_t - \Pi' x_t)] \end{aligned}$$

The joint density conditional on the first p observations can be written as

$$\begin{aligned} & f(W_T, W_{T-1}, W_{T-2}....W_1 | W_0, ., W_{-p+1}; \Theta) \\ = \quad & \prod_{t=1}^{T} f(W_t | W_{t-1}, W_{t-2}...., W_{-p+1}; \Theta) \end{aligned}$$

and taking logs,

$$\begin{aligned} L(\Theta) &= \sum_{t=1}^{T} log[f(W_t|W_{t-1}, W_{t-2}...., W_{-p+1}; \Theta)] \\ &= -(Tn/2)log(2\pi) + (T/2)log|\Omega^{-1}| - (1/2)\sum_{t=1}^{T}(W_t - \Pi' x_t)'\Omega^{-1}(W_t - \Pi' x_t) \end{aligned}$$

It turns out to be (for a proof see textbook) that the maximum likelihood estimator is

$$\hat{\Pi}' = [\sum_{t=1}^{T} W_t x_t'][\sum_{t=1}^{T} x_t x_t']^{-1}$$

where the j column is just

$$\hat{\pi}_{j\ (1\times \mathrm{np})} = [\sum_{t=1}^{T} W_{jt} x_t'][\sum_{t=1}^{T} x_t x_t']^{-1}$$

The Maximum likelihood estimator of Ω

We can now "concentrate" the likelihood using the previous results to find the MLE estimator of Ω (evaluated at the estimate of Π).

$$L(\Omega, \hat{\Pi}) = -(Tn/2)log(2\pi) + (T/2)log|\Omega^{-1}| - (1/2)\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}\hat{\varepsilon}_t$$

taking the derivative with respect to Ω^{-1}

$$\partial L(\Omega, \hat{\Pi})/\partial\Omega^{-1} = (T/2)\frac{\partial log|\Omega^{-1}|}{\partial\Omega^{-1}} - (1/2)\sum_{t=1}^{T}\frac{\partial\hat{\varepsilon}_t'\Omega^{-1}\hat{\varepsilon}_t}{\partial\Omega^{-1}}$$

and using the following results from matrix algebra: $\frac{\partial(x'Ax)}{\partial A} = xx'$ and $\frac{\partial log|A|}{\partial A} = (A')^{-1}$,we can differentiate the concentrated likelihood with respect to Ω^{-1}

$$\partial L(\Omega, \hat{\Pi})/\partial\Omega^{-1} = (T/2)\Omega' - (1/2)\sum_{t=1}^{T}(\hat{\varepsilon}_t\hat{\varepsilon}_t').$$

Equating this expression to zero we obtain the MLE of the variance-covariance matrix.

$$\hat{\Omega}' = (1/T)\sum_{t=1}^{T}(\hat{\varepsilon}_t\hat{\varepsilon}_t')$$

A very important result is that the row i, column i of $\hat{\Omega}$ is given by $\hat{\sigma}_i^2 = (1/T)\sum_{t=1}^{T}(\hat{\varepsilon}_{it}^2)$ which is just the average squared residual from a regression of a variable of the VAR on the p lags of all variables. Analogously the row i column

j of $\hat{\Omega}'$ is given by $\hat{\sigma}_{ij}^2 = (1/T)\sum_{t=1}^{T}(\hat{\varepsilon}_{it}\hat{\varepsilon}_{jt})$ which is the average product of the OLS residual for variable i and the OLS residual foe variable j. Therefore I can use OLS results to construct both $\hat{\Omega}$ and $\hat{\Pi}$.

How to choose the order of a VAR

The results of any test that we carry out using a VAR crucially depend on identifying correctly the order of that VAR. An easy way to attempt to identify the order of a VAR is to perform likelihood ratio tests. To do this turns out to be computationally very simple since the test can be constructed using OLS results.

Consider the likelihood function at is Maximum value of a VAR with p_0 lags, denoted

$$L_0(\hat{\Omega}, \hat{\Pi}) = -(Tn/2)log(2\pi) + (T/2)log|\hat{\Omega}_0^{-1}| - (1/2)\sum_{t=1}^{T}\hat{\varepsilon}_t'\hat{\Omega}_0^{-1}\hat{\varepsilon}_t.$$

Consider now the last term of this equation,

$$\begin{aligned}
(1/2)\sum_{t=1}^{T}\hat{\varepsilon}_t'\hat{\Omega}_0^{-1}\hat{\varepsilon}_t(ascalar) &= TR((1/2)\sum_{t=1}^{T}\hat{\varepsilon}_t'\hat{\Omega}_0^{-1}\hat{\varepsilon}_t) \\
&= (1/2)TR(\sum_{t=1}^{T}\hat{\Omega}_0^{-1}\hat{\varepsilon}_t\hat{\varepsilon}_t') \\
(since\ TR(A.B) &= TR(B.A)) \\
&= (1/2)TR(\hat{\Omega}_0^{-1}T\hat{\Omega}_0) \\
(since\hat{\Omega}_0 &= \sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t'/T) \\
&= (T/2)TR(I) \\
&= (nT)/2.
\end{aligned}$$

then

$$L_0(\hat{\Omega}, \hat{\Pi}) = -(Tn/2)log(2\pi) + (T/2)log|\hat{\Omega}_0^{-1}| - (nT)/2$$

If we want to test the Hypothesis that the VAR has p lags against p_0 lags we calculate the likelihood for the VAR with p_1 lags ($p_1 > p_0$)

$$L_1(\hat{\Omega}, \hat{\Pi}) = -(Tn/2)log(2\pi) + (T/2)log|\hat{\Omega}_1^{-1}| - (nT)/2$$

and compute the likelihood ratio which is

$$2(L_1(\hat{\Omega}, \hat{\Pi}) - L_0(\hat{\Omega}, \hat{\Pi})) = T(\log|\hat{\Omega}_1^{-1}| - \log|\hat{\Omega}_0^{-1}|)$$

which is distributed under the Null χ^2 with degrees of freedom equal to the number of restrictions imposed under H_0, $n^2(p_1 - p_0)$.

Sims (1980) proposed a modification of the likelihood ratio test to take into account small sample bias

$$(T-k)(\log|\hat{\Omega}_1^{-1}| - \log|\hat{\Omega}_0^{-1}|)$$

where $k = np_1$ = number of parameters estimated per equation.

Goodness of fit Criteria

The goodness of fit criteria are measures of how good a model is relative to others. They reflect a balance between the model's goodness of fit and the complexity of the model.

Typically, we want to minimize a scalar measure such as

$$C(p) = -2max(logL) + \beta(number\ of\ freely\ estimated\ parameters)$$

For Gaussian models, the maximized log-likelihood is proportional to

$$-(T/2)log|\Omega| \qquad\qquad (\text{since } |\Omega^{-1}| = 1/|\Omega| \ )$$

Hence, we choose p to minimize:

$$C(p) = Tlog|\Omega| + \beta(n^2 p)$$

AIC	$\beta = 2$ (Akaike information criterion)
SBC	$\beta = log(T)$ (Scharz Bayesian criterion)
HQ	$\beta = 2log(log(T))$ (Hannan-Quin criterion)

Alternatively the Akaike's prediction error (FPE) criterion chooses p so that to minimize the expected one -step ahead squared forecast error:

$$FPE = [\frac{T + np + 1}{T - np - 1}]^n |\Omega|$$

AIC and FPE are not consistent, so that asymptotically they overestimate p with positive probability. SBC and HQ are consistent in the sense that $\hat{p} \rightarrow p$.

Asymptotic Distribution of $\hat{\Pi}$

The maximum likelihood estimates of $\hat{\mathbf{\Pi}}$ and $\hat{\Omega}$ will give consistent estimates of the population parameters. Standard errors for $\hat{\mathbf{\Pi}}$ can be based on the usual OLS formulas.

Let $\hat{\pi}_T = vec(\hat{\mathbf{\Pi}}_T)$ denote the $nk_{\times}1$ vector of coefficients resulting from OLS regressions of each of the elements of W_t on x_t

Then

$$\sqrt{T}(\hat{\pi}_T - \pi) \underrightarrow{L} N(0, \Omega \otimes Q^{-1})$$

where $Q = E(x_t x_t')$.

This establishes that the standard OLS t and F statistics applied to the coefficients of any single equation in the VAR are asymptotically valid.

$$\sqrt{T}(\hat{\pi}_{iT} - \pi_i) \underrightarrow{L} N(0, (\sigma_i^2 Q^{-1}))$$

where $\sigma_i^2 = E(\varepsilon_{it}^2)$.

Main uses of Vector Autoregressions

i) Forecasting
ii) Testing Hypothesis
iii) Granger Causality
iv) Use of Impulse Response Functions
v) Use of the variance decomposition.

ii) Testing Rational expectations Hypothesis

The VAR methodology is very useful to test linear rational expectations hypothesis. These models usually impose non-linear cross equation restrictions between the parameters of the model which are tested using a likelihood ratio test which is distributed under the Null (that the model is correct) as a χ^2 distribution with degrees of freedom equal to the number of restrictions imposed by the model.

Consider a first order bivariate VAR

$$\begin{aligned} x_t &= a_{11}x_{t-1} + a_{12}y_{t-1} + \varepsilon_t \\ y_t &= a_{21}x_{t-1} + a_{22}y_{t-1} + \nu_t \end{aligned}$$

Assume that x_t is the interest rates differential, say $(i_t - i_t^*)$, and that y_t is the first difference of the logs of the spot exchange rate $(e_t - e_{t-1})$;

Then uncovered interest parity can be written as

$$x_t = E_t y_{t+1}.$$

Then if we condition on both sides of the previous equation on information available at $t-1$ we get the following set of non linear restrictions.

$$\begin{bmatrix} 1 & 0 \end{bmatrix} A = \begin{bmatrix} 0 & 1 \end{bmatrix} A^2$$

The above equation can be easily solved and yields the following nonlinear restrictions on the parameters

$$a_{11} = a_{22}a_{21}/(1-a_{21}) \qquad\qquad a_{12} = a_{22}^2/(1-a_{21})$$

Then we simply estimate the unrestricted model and the restricted (a function of only two parameters (and the variance-covariance)), and perform a likelihood ratio test, where $2(L_u - L_r)$~asymptotically χ^2 (number of restrictions=2).

iii)Granger Causality

One of the key questions that can be addressed with vector autoregressions is how useful some variables are for forecasting others.

Definition

The question investigated is whether a scalar y can help forecast another scalar x. If it cannot, then we say that y does not Granger-cause x

Then, y fails to Granger-cause x if for all $s > 0$ the mean squared error of a forecast of x_{t+s} based on $(x_t, x_{t-1}, ...)$ is the same as the MSE of a forecast of x_{t+s} based on $(x_t, x_{t-1}, ...)$ and $(y_t, y_{t-1}, ...)$. For linear functions

$$MSE[E(x_{t+s}|x_t, x_{t-1}, ...)] = MSE[E(x_{t+s}|x_t, x_{t-1}, ..., y_t, y_{t-1}, ...)]$$

Granger's reason for proposing this definition was that if an event Y is the cause of another event X, then the event Y should precede the event X.

Testing for Granger causality

Ho) y does not cause x or $a_{12} = 0$

We just regress both the general model

$$\text{i) } x_t = a_{11}x_{t-1} + a_{12}y_{t-1} + \varepsilon_{1t}$$

and the restricted model

$$\text{ii) } x_t = a_{11}x_{t-1} + \varepsilon'_{1t}$$

and compare the residuals sum squares $T(RRS(\varepsilon') - RRS(\varepsilon))/RRS(\varepsilon)$~$\chi^2(1)$ (asymptotically)

Granger-Causality Tests and Forward-Looking Behaviour

Let us assume risk neutral agents such that stock prices may be written as

$$P_t = \sum_{i=1}^{\infty} (1/(1+r))^i E(D_{t+i}|I_t)$$

suppose

$$D_t = d + u_t + \delta u_{t-1} + \nu_t$$

where u_t and ν_t are independent white noise processes, then

$$E_t D_{t+i} = \begin{cases} d + \delta u_t & \text{for } i = 1 \\ d & \text{for } i = 2, 3, ... \end{cases}$$

The stock prices will be given by

$$P_t = d/r + \delta u_t/(1+r)$$

Thus for this example the stock price is a white noise and could not be forecast on the basis of lagged prices or dividends. No series should granger cause stock prices.

Nevertheless, notice that using the stock price equation and rearranging terms, I might express

$$\delta u_{t-1} = (1+r)P_{t-1} - (1+r)d/r$$

Then substituting back in the Dividend process we get the following expression for dividends

$$D_t = d + u_t + (1+r)P_{t-1} - (1+r)d/r + \nu_t$$

Thus stock prices Granger cause dividends

The bivariate VAR takes the form

$$\begin{bmatrix} P_t \\ D_t \end{bmatrix} = \begin{bmatrix} d/r \\ -d/r \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ (1+r) & 0 \end{bmatrix} \begin{bmatrix} P_{t-1} \\ D_{t-1} \end{bmatrix} + \begin{bmatrix} \delta u_t/(1+r) \\ u_t + \nu_t \end{bmatrix}$$

Hence in this model, Granger causation runs in the opposite direction from the true causation. Dividends fail to G-C prices even though investors' perceptions of dividends are the sole determinant of stock prices. On the other hand, prices do "granger-cause" dividends, even though the market's evaluation of the stock in reality has no effect on the dividend process.

iv) The Impulse - Response Function

If a VAR is stationary it can always be written as an infinite vector moving average. Consider the following vector infinite moving average representation of W_t

$$W_t = \sum_{z=0}^{\infty} \psi_z \varepsilon_{t-z} \quad , \psi_0 = I$$

Analogously, if we lead the above expression s periods we get

$$W_{t+s} = \sum_{z=0}^{\infty} \psi_z \varepsilon_{t+s-z}$$

Therefore we can easily see from the above expression (evaluated at $z = s$) that matrix ψ_s has the interpretation of a dynamic multiplier

$$\psi_s = \frac{\partial W_{t+s}}{\partial \varepsilon_t'}$$

(dynamic multiplier or impulse response) where $(\psi_s)_{ij}$ = effect of a one unit increase in the j^{th} variable's innovation at time t (ε_{jt}) for the value of the i^{th} variable at time $t + s$ ($W_{i,t+s}$), holding all other innovations at all dates constant[2].

A simple way of finding these multipliers numerically is by simulation. To implement the simulation set $W_t = ... = W_{t-p} = 0$, then set $\varepsilon_{jt} = 1$ and all the other terms to zero, and simulate the system

$$W_t = A_1 W_{t-1} + A_2 W_{t-2} + ... + A_p W_{t-p} + \varepsilon_t$$

for $t, t + 1, t + s$, with $\varepsilon_{t+1}, \varepsilon_{t+2}, ... = 0$ This simulation corresponds to the J column of the matrix ψ_s . By doing this for other values of j we get the whole matrix.

A plot of $(\psi_s)_{ij}$, that is row i column j of ψ_s , as a function of s is called the *impulse response function*. It describes the response of $W_{i,t+s}$ to a one time impulse in W_{jt} with all other variables dated t or earlier held constant.

[2] Consider a first order two-variable Var. The relevant term of the infinite moving average is

$$\psi_s \varepsilon_t = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

Now assume a unit increase in ε_1 (i.e. $\varepsilon_1 = 1$ and $\varepsilon_2 = 0$), then the effect on W_{t+s} is $\begin{bmatrix} a \\ c \end{bmatrix}$. (that is **a** on the first variable and **c** on the second). Now varying s we can get a plot of the effects of a shock in the innovation of variable 1 on both variables. A similar argument can be constructed for a schock in ε_2, therefor the columns, first or second, of ψ_s represent the effect on each variable of W at time $t + s$ of a unit shock to the first or second variable innovation, keeping the other constant.

We can also define the Interim multipliers, which are given by the accumulated responses over m periods

$$\sum_{j=1}^{m} \psi_j,$$

and the long run multiplier which give the total accumulated effects for all future time periods:

$$\sum_{j=1}^{\infty} \psi_j.$$

The assumption that a shock in one innovation does not affect others is problematic since $E(\varepsilon_t \varepsilon_t') = \Omega \neq$ a diagonal matrix . This means that a shock in one variable is likely to be accompanied by a shock in another variable in the same period.

Since Ω is symmetric and positive definite, it can be expressed as $\Omega = ADA'$, where A is a lower triangular matrix and D is a diagonal Matrix.

Let $u_t = A^{-1}\varepsilon_t$, then

$$W_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} = \sum_{j=0}^{\infty} \psi_j A A^{-1} \varepsilon_{t-j} = \sum_{j=0}^{\infty} \psi_j^* u_{t-j}$$

where

$$\psi_j^* = \psi_j A$$

$$E(u_t u_t') = E(A^{-1}\varepsilon_t \varepsilon_t' (A^{-1})') = A^{-1}\Omega(A^{-1})' = A^{-1}ADA'(A^{-1})' = D$$

The matrix D gives the variance of u_{jt}

A plot of ψ_s^* as a function of s is known as an orthogonalized impulse response function.

The matrix

$$\psi_s^* = \frac{\partial W_{t+s}}{\partial u_t'}$$

gives the consequences of an increase in W_{jt} by a unit impulse in u_t .

In the new MA representation, it is reasonable to assume that a change in one component of u_t has no effect on the other components because the components are orthogonal.

Notice that $\psi_0^* = \psi_0 A = IA$ is lower triangular. This implies that the ordering of variables is of importance. The ordering has to be such that W_{lt} is the only one with a potential immediate impact on all other variables. W_{2t} may have an immediate impact on the last $n-2$ components but not on W_{1t} , and so on. The ordering cannot be determined with statistical methods.

v) Variance Decomposition

Consider the error in forecasting a VAR s periods ahead,

$$W_{t+S} - \widehat{W}_{t+S|t} = \sum_{j=0}^{s-1} \psi_j \varepsilon_{t+s-j}, \qquad \psi_0 = I$$

The mean squared error of this s-period ahead forecast is thus

$$MSE(\widehat{W}_{t+s|t}) = \Omega + \psi_1 \Omega \psi_1' + \psi_2 \Omega \psi_2' ..... + \psi_{s-1} \Omega \psi_{s-1}'$$

Let us now consider how each of the orthogonalized disturbances $(u_{1t}, ..u_{\rm nt})$ contributes to this MSE.

Write

$$\varepsilon_t = Au_t = a_1 u_{1t} + ...a_n u_{nt},$$

where a_j denotes the j^{th} column of the matrix A.

Recalling that the u's are uncorrelated, we get

$$\Omega = a_1 a_{1(n\times n)}' Var(u_{1t}) + a_2 a_{2(n\times n)}' Var(u_{2t}) + ...a_n a_{n(n\times n)}' Var(u_{\rm nt})$$

Substituting this in the MSE of the s period ahead forecast we get

$$MSE(\widehat{W}_{t+s|t}) = \sum_{j=1}^{n} Var(u_{jt})(a_j a_j' + \psi_1 a_j a_j' \psi_1' + \psi_2 a_j a_j' \psi_2' ... + \psi_{s-1} a_j a_j' \psi_{s-1}')$$

With this expression we can calculate the contribution of the j^{th} orthogonalized innovation to the MSE of the s-period ahead forecast.

$$Var(u_{jt})(a_j a_j' + \psi_1 a_j a_j' \psi_1' + \psi_2 a_j a_j' \psi_2' ... + \psi_{s-1} a_j a_j' \psi_{s-1}')$$

Again the magnitude in general depends on the ordering of the variables

Structural VAR's

Blanchard (1989)

To introduce structural information in a VAR there are several ways to proceed. Probably the most popular is to try to impose restrictions in the covariance matrix. For a VAR(p), there are $p(p+1)/2$ elements in the covariance matrix of the elements of p so we might consider models of the innovations of the form

$C\varepsilon_t = Du_t$ where the u_t's are uncorrelated with variances $\sigma_1^2, ..., \sigma_p^2$. Then, we can have $p(p+1)/2$ - p unknown terms in C and D.

How exactly one would want to specify C and D is left to structural reasoning, for example Blanchard (1989) considers the following structure

$$\varepsilon_{1t} = eu_{2t} + u_{1t}$$
$$\varepsilon_{2t} = c_{21}\varepsilon_{1t} + u_{2t}$$

Where u_{1t} and u_{2t} are regarded as demand and supply shocks, while ε_{1t} and ε_{2t} are output and unemployment innovations respectively. If $e = 0$ output just respond to demand shocks.

Blanchard and Quah (1989).

Probably the most well known approach is Blanchard and Quah (1989). They have a bivariate system with demand and supply shocks but they do not impose the assumption that the shocks are uncorrelated. Rather they argue that a demand shock should have a zero long-run effect while a supply shock will not. Hence they will have

$$\varepsilon_{1t} = a_1 u_{2t} + u_{1t}$$
$$\varepsilon_{2t} = a_2 u_{1t} + u_{2t}$$

where the covariance of u_{jt} is assumed to be zero.

To see how it works consider

$$W_t = A_1 W_{t-1} + A_2 W_{t-2} + ... + A_p W_{t-p} + \varepsilon_t$$

is estimated and the implied MA representation is

$$W_t = \sum_{z=0}^{\infty} \psi_z \varepsilon_{t-z} \qquad , \psi_0 = I.$$

In terms of the shocks of interest, we will write $\varepsilon_t = Au_t$ where A is now defined as $A = \begin{bmatrix} 1 & a_1 \\ a_2 & 1 \end{bmatrix}$.

Then the moving average representation becomes,

$$W_t = \sum_{z=0}^{\infty} \psi_z A u_{t-z} \qquad , \psi_0 = I$$

But we want the long run effect of a demand shock, taken to be u_{1t} , upon output say, W_{1t} , to be zero. The long run effect of u_{1t} on W_{1t} is just obtained by summing

$$[\sum_{z=0}^{\infty} \psi_z A]_{[1,1]} = 0$$

Let the first row of $\sum_{z=0}^{\infty} \psi_z$, be $[\delta_1, \delta_2]$, then the restriction is just $\delta_1 + a_2\delta_2 = 0$ or $a_2 = -\delta_1/\delta_2$.

Thus one parameter can be found from this restriction and other three come from the fact that

$V(\varepsilon_t) = A \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} A'$, since there are three unknowns in $V(\varepsilon_t)$ to determine a_1, σ_1^2 and σ_2^2. Then all that is needed is to estimate δ_1, δ_2 . Now since we know that

i) $\sum_{z=0}^{\infty} \psi_z = \psi(1)$,

ii) $\sum_{z=0}^{\infty} \psi_z L^i = \psi(L)$,

iii) $\sum_{z=0}^{\infty} \psi_z L^i = (I - A_1 L - A_2 L^2 - ... + A_p L^p)^{-1} = A(L)^{-1}$,

iv) $\psi(1) = (I - A_1 - A_2 - ..... + A_p)^{-1} = A(1)^{-1}$, then all the information that is needed for the impulse response function, is obtained from the estimated parameters in the VAR.

Non-Stationary Series.

There are three important types of time series which one is likely to find in financial econometrics: stationary (I(0)), trend stationary and non-stationary (I(1)).

A Stationary process

A (weakly) stationary time series has a constant mean, a constant variance and the covariance is independent of time. Stationarity is essential for standard econometric theory. Without it we cannot obtain consistent estimators.

A quick way of telling if a process is stationary is to plot the series against time. If the graph crosses the mean of the sample many times, chances are that the variable is stationary, otherwise that is an indication of persistent trends away from the mean of the series.

A Trend Stationary

A trend stationary variable is a variable whose mean grows around a fixed trend. This provides a classical way of describing an economic time series which grows at a constant rate. A trend-stationary series tends to evolve around a steady, upward sloping curve without big swings away from that curve. Detrending the series will give a stationary process. For simplicity assume that the following process.

$$y_t = \alpha + \mu t + \varepsilon_t \ where \ \varepsilon_t \tilde{} N(0, \sigma^2)$$

Notice that the mean of this process varies with time but the variance is constant.

$$E(y_t) = \alpha + \mu t$$

$$V(y_t) = E(\alpha + \mu t + \varepsilon_t - (\alpha + \mu t))^2 = \sigma^2$$

Notice that if you define a new variable, say $y_t^*, y_t^* = y_t - (\alpha + \mu t)$ then y_t is stationary.

A Non stationary Series: I(1) Processes

An autoregressive process of order p, AR(p), has a **unit root** if the polynomial in L, $(1 - \phi_1 L - ...\phi_p L^p)$ has a root equal to one. The simplest example of a process with a unit root is a random walk, i.e.,

$$y_t = y_{t-1} + \varepsilon_t \tag{1}$$

where ε_t is i.i.d. with zero mean and constant variance.

We can easily see that the variance of this processes does not exist: lagging the process one period we can write

$$y_{t-1} = y_{t-2} + \varepsilon_{t-1},$$

and substituting back in equation (1) we get

$$y_t = y_{t-2} + \varepsilon_{t-1} + \varepsilon_t.$$

Then, repeating this procedure we can easily show that

$$y_t = y_0 + \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{t-1} + \varepsilon_t$$

Then we can calculate the mean and the variance of this process:

The mean can be calculated assuming that y_0 is fixed, then the mean is constant over time

$$E(y_t) = E(y_0 + \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{t-1} + \varepsilon_t) = y_0$$

The variance of y_t , "conditional" on knowing y_0 , can be computed as

$$\begin{aligned} V(y_t) &= V(y_0 + \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{t-1} + \varepsilon_t) \\ &= V(\varepsilon_1) + V(\varepsilon_2) + ... + V(\varepsilon_{t-1}) + V(\varepsilon_t) = t\sigma^2 \end{aligned}$$

As we move further into the future this expression becomes infinite. We conclude that the variance of a unit root process is infinite.

A unit root process will only cross the mean of the sample very infrequently, and the process will experience long positive and negative strays away from the sample mean.

A process that has a unit root is also called **integrated of order one**, denoted as I(1). By contrast a stationary process is an **integrated of order zero** process, denoted as I(0).

Why is this Important?

The study of econometric models with non-stationary data has been one of the most important concerns of econometricians in the last 20 years. Therefore the topic is very vast and we just will *mention* some of the most important issues.

Spurious Regressions.

Granger and Newbold (1974) have shown that, using I(1), you can obtain an apparently significant regression (say with a high R^2) even if the regressor and the dependent variable are independent. He generated independent random

walks regress one against the other and obtain very high R^2 for this equation. They conclude that this result is spurious. As a rule of thumb whenever you obtain a very high R^2 and a very low DW you should suspect that the result are spurious.

Regressing Series that are integrated of the same order

Another important issue is that whenever you try to explain a variable, say y_t , by another variable, say x_t , you should check that these variables are integrated of the same order to obtain meaningful results.

How do we test for a unit root?

In this section we will show that the standard t-test cannot be applied for a process with a unit root. The standard testing procedure that we use for stationary series yields a degenerate distribution. We also find the distribution under these circumstances but; a) it turns out that is not a t-distribution and b) this distribution is biased to the left.

Consider the following model:

$$y_t = \alpha y_{t-1} + \varepsilon_t \tag{2}$$

where ε_t "is assumed" to be N(0,σ^2). It can easily be shown that asymptotically

$$\sqrt{T}(\hat{\alpha}_T - \alpha) \ \underrightarrow{L} \ N(0, (1-\alpha^2))$$

If we want to use this distribution for testing the null hypothesis that $\alpha = 1$, then we find that the distribution under the null "degenerates" (collapses in one point).

We find that the estimator converges in probability to the true parameter:

$$\sqrt{T}(\hat{\alpha}_T - \alpha) \ \underrightarrow{P} \ 0$$

Even though this is valid, it is not very useful for hypothesis testing.

To obtain a non-degenerate asymptotic distribution for $\hat{\alpha}_T$ in the unit root case, it turns out that we have to multiply by T and not by the square root of T. Then the unit root coefficient converges at a faster rate T than for the stationary case.

To get a better sense of why scaling by T is necessary when the true value of α is unity consider the OLS estimate

$$\hat{\alpha}_T = \frac{\sum_{t=1}^{T} y_{t-1} y_t}{\sum_{t=1}^{T} y_{t-1}^2}.$$

Then, substituting y_t by the AR(1) process we get that

$$\hat{\alpha}_T - \alpha = \frac{\sum_{t=1}^{T} y_{t-1}\varepsilon_t}{\sum_{t=1}^{T} y_{t-1}^2}$$

and multiplying in both sides by T, we get

$$T(\hat{\alpha}_T - \alpha) = \frac{T^{-1}\sum_{t=1}^{T} y_{t-1}\varepsilon_t}{T^{-2}\sum_{t=1}^{T} y_{t-1}^2}.$$

Now, under the null that $\alpha = 1$, y_t can be written as

$$\begin{aligned} y_t &= y_0 + \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{t-1} + \varepsilon_t \\ &= \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{t-1} + \varepsilon_t \qquad if\ we\ assume\ y_0 = 0. \end{aligned}$$

Then under the null that $\alpha = 1$, y_t ~N$(0, \sigma^2 t)$. To find the distribution of the *numerator* we need to do some easy but tedious algebra. We start by noting that under the null,

$$y_t^2 = (y_{t-1} + \varepsilon_t)^2 = y_{t-1}^2 + 2y_{t-1}\varepsilon_t + \varepsilon_t^2$$

and rearranging terms we obtain

$$y_{t-1}\varepsilon_t = \frac{1}{2}(y_t^2 - y_{t-1}^2 - \varepsilon_t^2).$$

Then, the sum which appears in the *numerator* can be expressed as

$$\sum_{t=1}^{T} y_{t-1}\varepsilon_t = \sum_{t=1}^{T} \frac{1}{2}(y_t^2 - y_{t-1}^2 - \varepsilon_t^2) = \frac{1}{2}(y_T^2 - y_0^2) - \sum_{t=1}^{T} \frac{1}{2}\varepsilon_t^2.$$

Then, recalling that $y_0 = 0$ and multiplying by (T^{-1}) we obtain the expression of the *numerator* as the sum of two terms

$$T^{-1}\sum_{t=1}^{T} y_{t-1}\varepsilon_t = (\frac{1}{2T})y_T^2 - \sum_{t=1}^{T}(\frac{1}{2T})\varepsilon_t^2$$

To find the distribution of this expression we divide each side by σ^2 which yields the following result

$$(\sigma^2 T)^{-1}\sum_{t=1}^{T} y_{t-1}\varepsilon_t = (1/2)(\frac{y_T}{\sigma\sqrt{T}})^2 - \sum_{t=1}^{T}(\frac{1}{2\sigma^2 T})\varepsilon_t^2$$

Consider the first term of this expression. Since we have shown above that y_t ~N$(0, \sigma^2 t)$, standardizing we obtain

$$(y_T/\sigma\sqrt{T})\tilde{}N(0,1),$$

and then squaring this expression we find that the first term of the numerator is distributed Chi-square

$$(y_T/\sigma\sqrt{T})^2 \tilde{} \chi^2(1).$$

It can be shown using the law of large numbers that the second term converges in probability to σ^2, i.e.

$$(1/T)\sum_{t=1}^{T}\varepsilon_t^2 \quad \underrightarrow{P} \quad \sigma^2, \qquad or \qquad (1/\sigma^2 T)\sum_{t=1}^{T}\varepsilon_t^2 \quad \underrightarrow{P} \quad 1$$

If we put both results together we can see that the numerator converges to

$$(\sigma^2 T)^{-1}\sum_{t=1}^{T} y_{t-1}\varepsilon_t \qquad \underrightarrow{L} \qquad (1/2)(X-1) \quad where \quad X \tilde{} \chi^2(1).$$

It can also be shown using the law of large numbers that the *denominator* converges in probability to

$$E(T^{-2}\sum_{t=1}^{T} y_{t-1}^2).$$

Now,as y_{t-1} ~N$(0, \sigma^2(t-1))$, then $E(y_{t-1}^2) = \sigma^2(t-1)$. Therefore the expected value of the denominator can be written as

$$E(T^{-2}\sum_{t=1}^{T} y_{t-1}^2) = T^{-2}\sigma^2\sum_{t=1}^{T}(t-1) = \sigma^2 T^{-2}(T-1)T/2$$

Then if we multiply $(\hat{\alpha}_T - \alpha)$ by T instead than by $\sqrt{T}$, we obtain a non-degenerate asymptotic distribution, but this distribution is not Gaussian.

For this very simple example it is rather easy to find the distribution using "standard" discrete time algebra. Unfortunately for many other processes is not as easy. Fortunately there is a convenient way of finding the distributions of processes with unit roots which relies in the is use of continuous time stochastic processes defined below.

Testing for the existence of unit roots

Consider the following model

$$y_t = \mu + \beta t + \alpha y_{t-1} + \varepsilon_t \tag{3}$$

where ε_t "is assumed" to be N(0,σ^2)

We want to test the Hypothesis of the existence of a unit root therefore we set the following null and alternative hypothesis.

H_0) $\alpha = 1$ (unit root)

H_1) $\alpha < 1$ (Integrated of order zero)

The obvious estimator of is the OLS estimator, $\hat{\beta}$. The problem is that under the null hypothesis there is considerable evidence of the non - adequacy of the asymptotic (approximate in large samples) distribution. Therefore

Equation (3) can be reparameterizied as

$$\Delta y_t = \mu + (\alpha - 1)y_{t-1} + \beta t + \varepsilon_t$$

or

$$\Delta y_t = \mu + \lambda y_{t-1} + \beta t + \varepsilon_t \tag{4}$$

For this expression the relevant hypothesis should be written as

$H_0)\lambda = 0$ (unit root)

$H_1)\lambda < 0$ (Integrated of order zero)

Fuller (1976) tabulated, using Monte Carlo methods, critical values for alternative cases, for example for a sample size of 100 the 5 % critical values are

$\mu = 0,\ \beta = 0$	-2.24
$\mu \neq 0,\ \beta = 0$	-3.17
$\mu \neq 0,\ \beta \neq 0$	-3.73

Therefore the method simply consist to check the t-statistic of $\hat{\lambda}$ against the critical values of Fuller (1976). Notice that the critical values depend on

i) the sample size

ii) whether you include a constant and/or a time trend.

This procedure is only valid when there is no evidence of serial correlation in the residuals, $\hat{\varepsilon}_t$. To see if this condition is satisfied you should look at the diagnostic tests for serial correlation in the regression. If there is serial correlation you should need to include additional lags, say $\Delta y_{t-1}, \Delta y_{t-2}$ etc. to equation (4) until the serial correlation of the residuals disappears, that is

Augmented Dickey Fuller

$$\Delta y_t = \mu + \lambda y_{t-1} + \beta t + \alpha_1 \Delta y_{t-1} + ... + \alpha_k \Delta y_{t-k} + \varepsilon_t$$

In this case we chose to augment the regression with k lags. This is usually denoted as ADF(k). To choose the order of augmentation of the DF regression several procedures have been proposed in the literature. Some of these consist in:

(*i*) *choosing k as a function of the number of observations as in Schwert (1989)*

$k = INT(12(T/100)^{1/12})$

(*ii*) *information based rules such as AIC and BIC.*

(*iii*) *Sequential rules*

General to specific seems to be preferable to the other methods.

Small sample properties of the Dickey Fuller tests.

The power (the ability to reject the null Hypothesis) of the Dickey Fuller Tests is notoriously weak. Thus it can be difficult to reject the null of a unit root test even if the true series is stationary. That is, most of the time an ADF test will not reject the null of unit root even if the true model is an autoregressive model with an autoregressive coefficient of say .8 (that is an Integrated of order zero series). In addition the ADF tests often come up with conflicting results depending in the order of the lag structure. A good practice is to start with a quite general model and delete the non significant lags of Δy_t.

Phillips-Perron-type tests for unit roots

The ADF test includes additional lagged terms to account for the fact that the DGP might be more complicated than an AR(1). An alternative approach is that suggested by Phillips(1987) and Perron (1988). They make a non-parametric correction to the standard deviation which provides a consistent estimator of the variance.

They use

$$S_{Tl}^2 = T^{-1}\sum_{t=1}^{T}(\varepsilon_t^2) + 2T^{-1}\sum_{t=1}^{l}\sum_{t=j+1}^{T}\varepsilon_t\varepsilon_{t-j}$$

$$S_{\varepsilon}^2 = T^{-1}\sum_{t=1}^{T}(\varepsilon_t^2)$$

where $_l$ is the lag truncation parameter used to ensure that the autocorrelation of the residuals is fully captured.

An asymptotically valid test $\phi = 1$, for

$$\Delta y_t = \mu + (\phi - 1)y_{t-1} + \varepsilon_t \quad \varepsilon_t \tilde{} iiD(0, \sigma^2)$$

when the underlying DGP is not necessarily an AR(1) process, is given by the Phillips Z-test.

$$Z(\tau_\mu) = (S_\varepsilon / S_{Tl})\tau_\mu - (1/2)(S_{Tl}^2 - S_\varepsilon^2)[S_{Tl}[T^2 \sum_{t=2}^{T}(y_t - \overline{y})^2]^{.5}]^{-1}$$

where τ_μ is the t-statistic associated with testing the null hypothesis $\rho = 1$. The critical values for this test statistic are the same as those used for the same case in the fuller table. Monte Carlo work suggests that the Phillips-type test has poor size properties (tendency to over reject when is true) when the underlying DGP has large negative MA components.

Structural breaks and unit roots

We have already mentioned that the ADF test has very low power to reject the null hypothesis of the existence of a unit root. Therefore is very difficult using this type of test to distinguish between an autoregressive process with root, say .95, and a unit root process. This is particularly true if there is a structural break in the mean of the series. Perron (1988) have shown that an I(0) process with a structural break in the mean will be difficult to distinguish from a I(1) process. If we know were the break takes place, the natural thing to do, is to partial out the break by using dummy variables and test for unit roots once the break has been partialled out.

A possible solution to try to identify these breaks is to perform the ADF test recursively and to compute recursive t-statistics.

Perron (1989) showed that if a series is stationary around a deterministic time trend which has undergone a permanent shift sometime during the period under consideration, failure to take account of this change in the slope will be mistaken by the usual ADF unit root test as a persistent innovation to a stochastic (non-stationary) trend. That is, a unit root test which does not take into account a break in the series will have very low power. There is a similar loss in power if there is a shift in the intercept.

If the breaks in the series are known then it is relatively simple to adjust the ADF test by including dummy variables to ensure there are as many deterministic regressors as there are deterministic components in the DGP.

However is unlikely that we will know the break then we can proceed by using the critical values provided by Banjeree, Lumsdaine and Stock (1992). They designed a testing strategy and tables which account for a random break to take palace in any place in the sample. Then they designed tables which are constructed under the null of a unit root but allowing for a random structural break in the sample. They proposed two tests, one based on the recursive computation of the t - statistics of the ADF regression and another based on the rolling computation. They compare the minimum values whith the critical values of tables generated under this scenarios

Recursive t-statistics

The recursive ADF -statistic is computed using sub samples $t = 1..k\ for\ k = k_0, ...T$, where k is the start up value and T is the sample size of the full sample. The most general model (with drift and trend) is estimated for each sub sample and the minimum value of $\tau_\tau(k/T)$ across all the sub samples is chosen and compared with the table provided by Banjeree, Lumsdaine and Stock

Rolling ADF tests

This method could also be applied using a (large enough) window (of size k) to see if there are clear changes in the pattern of a series. The most general model (with drift and trend) is estimated for each sub sample and the minimum value of $\tau_\tau(k/T)$ across all the sub samples is chosen and compared with the table.

T	Percentile	τ_τ	Recursive min τ_τ	Rolling min τ_τ
100	.025	-3.73	-4.62	-5.29
	.050	-3.45	-4.33	-5.01
	.100	-3.15	-4.00	-4.71
250	.025	-3.69	-4.42	-5.07
	.050	-3.43	-4.18	-4.85
	.100	-3.13	-3.91	-4.59
500	.025	-3.68	-4.42	-5.00
	.050	-3.42	-4.18	-4.79
	.100	-3.13	-3.88	-4.55

Tests with stationarity as null: The **KPSS** test

Consider the following model.

$$y_t = \alpha + \delta t + \xi_t + \varepsilon_t$$

where ε_t is a stationary process and ξ_t is a random walk given by

$$\xi_t = \xi_{t-1} + u_t \quad u_t \tilde{} iid(0, \sigma_u^2)$$

The null of stationarity is formulated as

$H_0) \sigma_u^2 = 0$

The test statistic for this hypothesis is given by

$$LM = \frac{\sum_{t=1}^{T} S_t^2}{\hat{\sigma}_e^2}$$

where e_t are the residuals of a regression of y_t on a constant and a time trend, $\hat{\sigma}_e^2$ is the residual variance for this regression and S_t is the partial sum of e_t defined by

$$S_t = \sum_{i=t}^{T} e_i \quad t = 1, 2, .., T.$$

For testing the null of the level stationary instead of trend stationary the test is constructed the same way except that e_t is obtained as the residual from a regression of y_t on an intercept only. The test is an upper tail test. When the errors are i.id. the asymptotic distribution of the test is derived in Nabeya and Tanaka (1988). In other cases need to be conveniently adjusted.

Variance Ratio Tests

Given the low power of the ADF test, the variance ratio test will provide us with another tool to discriminate between (trend) stationary and non-stationary series.

Consider y_t and assume that it follows follows a random walk, i.e.

$$y_t = y_{t-1} + \varepsilon_t$$

then by iterative substitution we know that

$$y_t = y_{t-k} + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \varepsilon_{t-3} + ... + \varepsilon_{t-k+1}$$

Now if we denote the difference between y_t and y_{t-k} as $\Delta_k y_t$, then

$$\Delta_k y_t = \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \varepsilon_{t-3} + ... + \varepsilon_{t-k+1}$$

and clearly the variance of $\Delta_k y_t$, is $\sigma^2 k$, where σ^2 is the variance of ε.

We can define a "variance ratio" function (a function of k) as

$$\lambda_1(k) = \frac{Var(\Delta_k y_t)}{Var(\Delta_1 y_t)} = k.$$

Therefore a plot of λ_1 against k should be an increasing straight line. Alternatively we may define a new function $\lambda_2(k)$ as $\lambda_2(k) = \lambda_1(k)/k$, then if there is a unit root, $\lambda_2(k)$ tends to one when k tends to infinite.

However if y_t *does not* contain a unit root it can be shown that the lim $\lambda_2(k)$ when k tends to infinite is equal to zero. To see this assume the following *AR(1)* process

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \qquad t = 1, ..., T$$

we have seen that by iterative substitution we can express this process as

$$y_t = \phi_1^k y_{t-k} + \phi_1^{k-1} \varepsilon_{t-(k-1)} + \phi_1^{k-2} \varepsilon_{t-(k-2)} + ... + \phi_1 \varepsilon_{t-1} + \varepsilon_t$$

Then subtracting y_{t-k} in both sides of the equation we get the following expression

$$y_t - y_{t-k} = (\phi_1^k - 1)y_{t-k} + \phi_1^{k-1}\varepsilon_{t-(k-1)} + \phi_1^{k-2}\varepsilon_{t-(k-2)} + ... + \phi_1\varepsilon_{t-1} + \varepsilon_t$$

Then the variance of $y_t - y_{t-k}$, $Var(\Delta_k y_t)$ is equal to

$$V(y_t - y_{t-k}) = (\phi_1^k - 1)^2 V(y_{t-k}) + V(\sum_{j=0}^{k-1} \phi_1^j \varepsilon_{t-j})$$

Notice that

$$V(y_{t-k}) = V(y_t) = (1/(1-\phi_1^2))\sigma^2$$

and

$$V(\sum_{j=0}^{k-1} \phi_1^j \varepsilon_{t-j}) = ((1-\phi_1^{2k})/(1-\phi_1^2))\sigma^2,$$

therefore, we can write the variance of Δy_{t-k} as

$$V(y_t - y_{t-k}) = (\phi_1^k - 1)^2(1/(1-\phi_1^2))\sigma^2 + ((1-\phi_1^{2k})/(1-\phi_1^2))\sigma^2$$

In the same way we can express for a stationary process the variance of the first difference of y_t, $(y_t - y_{t-1}) = (\phi_1 - 1)y_{t-1} + \varepsilon_t$.

$$V(y_t - y_{t-1}) = (\phi_1 - 1)^2 V(y_{t-1}) + \sigma^2 = (\phi_1 - 1)(1/(1-\phi_1^2))\sigma^2 + \sigma^2$$

then the variance ratio can be written as

$$\lambda_1(k) = \frac{(\phi_1^k - 1)^2(1/(1-\phi_1^2))\sigma^2 + (\ (1-\phi_1^{2k})/(1-\phi_1^2)\ )\sigma^2}{(\phi_1 - 1)^2(1/(1-\phi_1^2))\sigma^2 + \sigma^2}$$

then the limit of $\lambda_1(k)$ when k tends to infinite for a stationary process is

$$lim_{k\to\infty}\lambda_1(k) = \frac{1}{1-\phi_1}$$

which is a constant provided that $\phi_1 \neq 1$.

It should be clear from the previous result that the limit of $\lambda_2(k)$ equals 0 when k tends to infinite.

Trend stationary vs difference stationary processes

A trend stationary variable may be written as

$$y_t = \alpha + \mu t + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \theta_3\varepsilon_{t-3} + ...$$

then y_{t-k} is simply

$$y_{t-k} = \alpha + \mu(t-k) + \varepsilon_{t-k} + \theta_1\varepsilon_{t-1-k} + \theta_2\varepsilon_{t-2-k} + \theta_3\varepsilon_{t-3-k} + ...$$

The k^{th} difference can be obtained simply by subtracting the two above equations

$$\begin{aligned} y_t - y_{t-k} &= \mu k + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_{k-1}\varepsilon_{t-(k-1)} + (\theta_k - 1)\varepsilon_{t-k} + \\ &\quad (\theta_{k+1} - \theta_1)\varepsilon_{t-(K-1)} + ... + (\theta_{k+q} - \theta_q)\varepsilon_{t-q} + ... \\ &= \mu k + \sum_{j=0}^{k-1}\theta_j\varepsilon_{t-j} + \sum_{j=0}^{\infty}(\theta_{k+j} - \theta_j)^2\varepsilon_{t-j} \end{aligned}$$

Then the variance of $y_t - y_{t-k}$ may be written as

$$V(y_t - y_{t-k}) = \sigma^2\sum_{j=0}^{k-1}\theta_j^2 + \sigma^2\sum_{j=0}^{\infty}(\theta_{k+j} - \theta_j)^2$$

From the previous equation we can see that when k tends to infinity, the variance of $\Delta_k y_t$ is equal to

$$V(\Delta_k y_t) = 2\sigma^2\sum_{j=0}^{k-1}\theta_j^2$$

Now the first difference of a trend stationary process, Δy_t is

$$\begin{aligned} y_t - y_{t-1} &= \mu + \varepsilon_t + (\theta_1 - 1)\varepsilon_{t-1} + ... + (\theta_{k+1} - \theta_k)\varepsilon_{t-(k+1)} \\ &\quad +... + (\theta_{k+q} - \theta_{k+q-1})\varepsilon_{t-q} + ... \end{aligned}$$

Then the variance of the first difference can be written as

$$\begin{aligned} V(y_t - y_{t-1}) &= V(\varepsilon_t + (\theta_1 - 1)\varepsilon_{t-1} + ... + (\theta_{k+1} - \theta_k)\varepsilon_{t-(k+1)} \\ &\quad +... + (\theta_{k+q} - \theta_{k+q-1})\varepsilon_{t-(k+q)} + ...) \\ &= \sigma^2(1 + \sum_{j=0}^{\infty}(\theta_{j+1} - \theta_j)^2) \end{aligned}$$

The variance ratio should be

$$\lambda_1(k) = \frac{Var(\Delta_k y_t)}{Var(\Delta_1 y_t)} = \frac{\sum_{j=0}^{k-1}\theta_j^2 + \sigma^2\sum_{j=0}^{\infty}(\theta_{k+j} - \theta_j)^2}{(1 + \sum_{j=0}^{\infty}(\theta_{j+1} - \theta_j)^2)},$$

and the limit when k tends to infinity is

$$\lim_{k\to\to\infty} \lambda_1(k) = \lim_{k\to\to\infty} \frac{Var(\Delta_k y_t)}{Var(\Delta_1 y_t)} = \lim_{k\to\to\infty} \frac{2\sum_{j=0}^{k-1}\theta_j^2}{(1+\sum_{j=0}^{\infty}(\theta_{j+1}-\theta_j)^2)}.$$

This expression is a constant and might be greater or smaller than one depending on the θ_j values. Therefore can simply distinguish between the two models by simply noting that under the random walk assumption λ_1 increase with k and that under the trend stationary assumption λ_1 tends to a constant. Alternatively we can consider λ_2 $Var(\Delta_k y_t)/k$. and note both, that when the model is a random walk this expression tends to 1 (see proof above), and that this ratio should tend to zero when k tends to infinity since $Var(\Delta_k y_t)$ is constant for the trend stationary model.

Sampling distribution of λ(k) under the Random Walk Hypothesis.

$$H_0)\ \alpha = 1\ or \quad y_t = \mu + y_{t-1} + \varepsilon_t \qquad\qquad \varepsilon_t \tilde{} IIDN(0,)$$

It can be shown that asymptotically under the null, $\sqrt{Tk}(\hat{\lambda}_2(k) - 1) \overset{a}{\sim} N(0, 2(k-1))$. Then tests of the null Hypothesis can be carried out on the standardized statistics.

1 Appendix

Brownian Motion

To show what is a BM we may start first considering a simple RW without drift.

$$y_t = y_{t-1} + \varepsilon_t, \quad \varepsilon_t \tilde{} N(0,1)$$

or

$$y_t = \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{t-1} + \varepsilon_t \quad and \quad y_t \tilde{} N(0,t)$$

Now consider the change between t and s

$$y_s - y_t = \varepsilon_t + \varepsilon_{t+1} + ... + \varepsilon_s$$

is $\tilde{} N(0, s-t)$ and is independent of changes between dates r and q whenever $t < s < r < q$.

Consider now the change between y_t and y_{t-1} with $\varepsilon_t \tilde{} N(0,1)$, and suppose we view ε_t as the sum of two independent Gaussian variables.

$$\varepsilon_t = e_{1t} + e_{2t} \quad with \quad e_{it} \tilde{} N(0, 1/2)$$

We may associate e_{1t} with the change between y_{t-1} and y_t at some interim point, say $y_{t-1/2}$, such that

$$y_{t-1/2} - y_{t-1} = e_{1t}$$

and

$$y_t - y_{t-1/2} = e_{2t}$$

Sampled at an integer $y_t - y_{t-1}$ ~N(0,1), but we can consider n possible divisions as above such that,

$$y_t - y_{t-1} = e_{1t} + e_{2t} + ... + e_{Nt}$$

where e_{it} ~i.i.d N(0,1/N) .

The limit when $N \to \infty$ is a continuous process called a Standard Brownian Motion. The value this process takes at date t is denoted $W(t)$. A realization of a continuous time process can be viewed as a stochastic function, denoted $W(.)$ where $W : t \in [0, \infty) \to R$.

Definition of a SBM

$W(.)$ is a continuous time process,associating each date $t \in [0,1]$ with the scalar $W(t)$ such that

(a) $W(0) = 0$

(*b*) for any dates $0 < t_1 < t_2 ... < t_k < 1$,

$W(t_2) - W(t_1), ..., W(t_k) - W(t_{k-1})$.
are independent multivariate Gaussian with $W(t_s) - W(t_t)$~N(0,$s - t$)

(*c*) $W(t)$ is continuous with probability 1.

The functional Central Limit Theorem

Recall the simplest version of the Central limit Theorem.

If ε_t~i.i.d with mean zero and variance σ^2, then the sample mean $\tilde{\varepsilon}_T = T^{-1}\sum_{t=1}^{T}\varepsilon_t$ and the central limit theorem states that

$$\sqrt{T}\tilde{\varepsilon}_T \quad \underrightarrow{L} \quad N(0,\sigma^2)$$

Consider now an estimator based on the following principle: Given a sample size T, we calculate the mean of the first half of the sample and throw out the rest of the observations.

$$\widetilde{\varepsilon}_{[T/2]^*} = ([T/2]^*)^{-1}\sum_{t=1}^{[T/2]}\varepsilon_t$$

where $[T/2]^*$ is the larger integer $\geq$ than $T/2$, i.e.

$$[T/2]^* = T/2 \; for \; T \; even$$

$$[T/2]^* = (T-1)/2 \; for \; T \; odd.$$

This will also satisfy

$$\sqrt{[T/2]^*}\tilde{\varepsilon}_{[T/2]^*} \quad \underrightarrow{L} \quad N(0,\sigma^2)$$

Moreover the estimator will be independent of an estimator that uses only the second half of the sample.

More generally we can construct a Variable $X_T(r)$ from the sample mean of the first r^{th} fraction of observations, where $r \in [0,1]$ defined by

$$X_T(r) = \frac{1}{T}\sum_{t=1}^{Tr^*}\varepsilon_t$$

For any given realization $X_T(r)$ is a step function in r, with

$$X_T(r) = \begin{cases} 0 & \text{for } 0 \leq r < (1/T) \\ \varepsilon_1/T & \text{for } (1/T) \leq r < (2/T) \\ (\varepsilon_1+\varepsilon_2)/T & \text{for } (2/T) \leq r < (3/T) \\ (\varepsilon_1+\varepsilon_2+\varepsilon_3+\varepsilon_4+...+\varepsilon_T)/T & \text{for } r = 1 \end{cases}$$

then

$$\sqrt{T}X_T(r) = (1/\sqrt{T})\sum_{t=1}^{Tr^*}\varepsilon_t = (\sqrt{Tr^*}/\sqrt{T})(1/\sqrt{Tr^*})\sum_{t=1}^{Tr^*}\varepsilon_t$$

but

$$(1/\sqrt{Tr^*})\sum_{t=1}^{Tr^*}\varepsilon_t \quad \underrightarrow{L} \quad N(0,\sigma^2) \quad while \quad (\sqrt{Tr^*}/\sqrt{T}) \to \sqrt{r}$$

Hence the asymptotic distribution of $\sqrt{T}X(r)$ is that of $\sqrt{r}$ times a $N(0,\sigma^2)$ or $\sqrt{T}X_T(r) \underrightarrow{L} N(0,\sigma^2 r)$.

Clearly this implies that

$$\frac{\sqrt{T}X_T(r)}{\sigma} \quad \underrightarrow{L} \quad N(0,r).$$

Notice that this is evaluated at a given value r (NB $\frac{\sqrt{T}X_T(r)}{\sigma}$ is a random variable).

We could consider the behaviour of the sample mean based on Tr_1^* through Tr_2^* for $r_2 > r_1$ and conclude that

$$\frac{\sqrt{T}(X_T(r_2) - X_T(r_1))}{\sigma} \quad \underrightarrow{L} \quad N(0, r_2 - r_1)$$

Then a sequence of stochastic functions $\frac{\sqrt{T}(X_T(.)}{\sigma}|_{T=1}^{\infty}$ has an asymptotic probability law that is described by a standard Brownian motion.

$$\frac{\sqrt{T}(X_T(.))}{\sigma} \quad \underrightarrow{L} \quad W(.) \ (this\ is\ a\ random\ function)$$

This function evaluated at $r = 1$ is just the sample mean, $X_T(1) = T^{-1}\sum_{t=1}^{T}\varepsilon_t$, then when $r = 1$ the CLT is a special case of this function, that is $\frac{\sqrt{T}(X_T(1))}{\sigma} \quad \underrightarrow{L} \quad W(1) = N(0,1)$

Convergence in Functions

Let $S(.)$ represent a continuous time stochastic process with $S(r)$ representing its value at some date r for $r \in [0,1]$. Suppose further that for any given realization $S(.)$ is a continuous function of r with probability 1, being $\{S_T(.)\}|_{T=1}^{\infty}$ a sequence of such continuous functions,

We say that

$$S_T(.) \underrightarrow{L} S(.) \text{ whenever the following holds:}$$

a) For any finite collection of k particular dates $0 < r_1 < r_2 .... < r_k \leq 1$, the sequence of k dimensional random vectors $y_T \underrightarrow{L} y$ where

$$y_T \equiv \begin{bmatrix} S_T(r_1) \\ S_T(r_k) \end{bmatrix} \quad and \; y \equiv \begin{bmatrix} S(r_1) \\ S(r_k) \end{bmatrix}$$

b) For each $\varepsilon > 0$ the probability that $S_T(r_1)$ differs from $S_T(r_2)$ for any dates r_1 and r_2 within a distance δ of each other goes to zero uniformly in T as $\delta \to 0$..

c) $P\{|S_T(0)| > \lambda\} \to 0$ uniformly in T as $\lambda \to \infty$.

This definition applies to sequences of continuous functions though $X_T(r)$ is a discontinuous step function. Fortunately the discontinuities occur at countable points.

Convergence for a sequence of Random Functions.

It will be helpful to extend the earlier definition of convergence in probability to sequences of random functions.

Let $\{S_T(.)\}|_{T=1}^{\infty}$ and $\{V_T(.)\}|_{T=1}^{\infty}$ denote sequences of random continuous functions with

$$S_T : r \in [0,1] \to R, \quad V_T : r \in [0,1] \to R.$$

Let the scalar Y_T represent the largest amount by which $S_T(r)$ differs from $V_T(r)$.

$$Y_T = Sup|S_T(r) - V_T(r)|$$

Then $\{Y_T\}|_{T=1}^{\infty}$ is a sequence of random variables and we could talk about its probability limit. If the sequence converges in probability to 0, then we can say that

$$S_T(r) \quad \underrightarrow{P} \quad V_T(r) \quad or \quad Sup|S_T(r) - V_T(r)| \quad \underrightarrow{P} \quad 0.$$

This can be generalized to sequences of functions. For example, if $\{S_T(.)\}|_{T=1}^{\infty}$ and $\{V_T(.)\}|_{T=1}^{\infty}$ are sequences of continuous functions with $V_T(.) \underrightarrow{P} S_T(.)$ and $S_T(.) \underrightarrow{L} S(.)$, for $S(.)$ a continuous function, then $V_T(.) \underrightarrow{L} S(.)$

Continuous Mapping Theorem

If $g(.)$ is a continuous functional, which could associate a real variable y with the stochastic function $S(.)$, then the theorem states that if $S_T(.) \underrightarrow{L} S(.)$ and $g(.)$ is a continuous functional, then $g(S_T(.)) \underrightarrow{L} g(S(.))$

(Examples of continuous functionals)

$y = \int_0^1 S(r)dr$ or $y = \int_0^1 [S(r)]^2 dr$

Example

Given $\frac{\sqrt{T}X_T(.)}{\sigma} \underrightarrow{L} W(.)$, we can simply use the theorem to get

$$\sqrt{T}X_T(.) \quad \underrightarrow{L} \quad \sigma W(.)$$

or

$$[\sqrt{T}X_T(.)]^2 \quad \underrightarrow{L} \quad [\sigma W(.)]^2$$

Applications to Unit Root Processes

Example: a Random Walk.

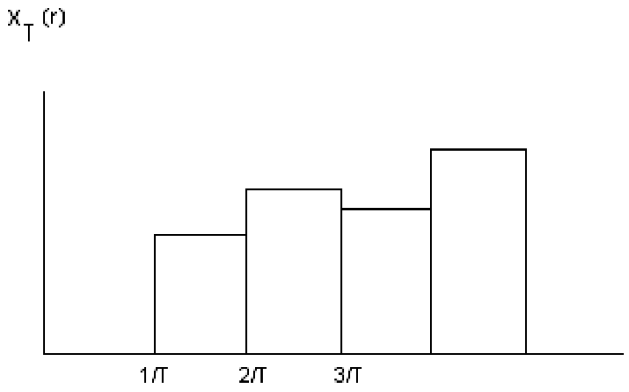
$$y_t = y_{t-1} + \varepsilon_t \qquad \varepsilon_t \tilde{} iid(0, \sigma^2)$$

or

$$y_t = \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{t-1} + \varepsilon_t$$

this can be used to express the stochastic function

$$X_T(r) = \begin{cases} 0 & \text{for } 0 \leq r < (1/T) \\ y_1/T & \text{for } (1/T) \leq r < (2/T) \\ y_2/T & \text{for } (2/T) \leq r < (3/T) \\ y_T/T & \text{for } r = 1 \end{cases}$$



Notice that the t^{th} rectangle has width $1/T$ and height y_{t-1}/T. The total area is then

$$\int_0^1 X_T(r)dr = y_1/T^2 + y_2/T^2 + ... + y_{T-1}/T^2$$

Multiplying both sides by $\sqrt{T}$

$$\int_0^1 \sqrt{T}X_T(r)dr = T^{-3/2}\sum_{t=1}^{T} y_{t-1}$$

But we know from the continuous mapping theorem that as $T \to \infty$

$$\int_0^1 \sqrt{T}X_T(r)dr \quad \underrightarrow{L} \quad \int_0^1 [\sigma W(r)]dr$$

(since $[\sqrt{T}X_T(.)] \quad \underrightarrow{L} \quad [\sigma W(.)]$), implying

$$T^{-3/2}\sum_{t=1}^{T} y_{t-1} \quad \underrightarrow{L} \quad \int_0^1 [\sigma W(r)]dr$$

Distribution of the OLS estimator under the Unit root Hypothesis

Consider now the following autoregressive process

$$y_t = \rho y_{t-1} + \varepsilon_t \quad \varepsilon_t \tilde{} iid(0, \sigma^2)$$

$$(\hat{\rho}_T - \rho) = \frac{\sum_{t=1}^{T} y_{t-1}\varepsilon_t}{\sum_{t=1}^{T} y_{t-1}^2}$$

then and assume that we want to test the null

$H_0)$ $\rho = 1$.

$$T(\hat{\rho}_T - 1) = \frac{T^{-1}\sum_{t=1}^{T} y_{t-1}\varepsilon_t}{T^{-2}\sum_{t=1}^{T} y_{t-1}^2}$$

Then the distribution of the denominator can be easily obtained since

$$T^{-2}\sum_{t=1}^{T} y_{t-1}^2 \quad \underrightarrow{L} \quad \int_0^1 [\sigma W(r)]^2 dr$$

because

$$[\sqrt{T}X_T(.)] \underrightarrow{L} [\sigma W(.)],$$

$$[\sqrt{T}X_T(.)]^2 \underrightarrow{L} [\sigma W(.)]^2,$$

$$\int_0^1 [\sqrt{T}X_T(r)]^2 dr \quad \underrightarrow{L} \quad \int_0^1 [\sigma W(r)]^2 dr,$$

and

$$T^{-2}\sum_{t=1}^{T} y_{t-1}^2 = \int_0^1 [\sqrt{T}X_T(r)]^2 dr$$

This is because

$$(X_T(r))^2 = \begin{cases} 0 & \text{for } 0 \le r < (1/T) \\ (y_1/T)^2 & \text{for } (1/T) \le r < (2/T) \\ (y_2/T)^2 & \text{for } (2/T) \le r < (3/T) \\ . & . \\ (y_T/T)^2 & \text{for } r = 1 \end{cases}$$

and

$$\int_0^1 [\sqrt{T}X_T(r)]^2 dr = T\left((y_1)^2/T^3 + (y_2)^2/T^3 + ... + (y_{T-1})^2/T^3 \right).$$

Also

$$T^{-1}\sum_{t=1}^{T} y_{t-1}\varepsilon_t \quad \underrightarrow{L} \quad (1/2)\sigma^2[W(1)^2 - 1]$$

(Recall that $T^{-1}\sum_{t=1}^{T} y_{t-1}\varepsilon_t = (1/2T)y_T^2 - \sum_{t=1}^{T}(1/2T)\varepsilon_t^2$ (when $y_0 = 0$)) then

$$\boxed{T(\hat{\rho}_T - 1) = \frac{(1/2)\sigma^2[W(1)^2 - 1]}{\int_0^1 [\sigma W(r)]^2 dr}}$$

Recall that $W(1)^2$ is chi square with one degree of freedom. The probability that a Chi-Square is less than 1 is .68 and since the denominator is positive, the probability that $\hat{\rho}_T - 1$ is negative approaches 0.68 as T tends to infinity. In other words in two thirds of the samples generated by a RW, the estimate will be less than the true value of unity or negative values will be twice as likely than positive values.

In practice critical values for the random variable $T(\hat{\rho}_T - 1)$ are found by calculating the exact small-sample distribution assuming the innovations are Gaussian, generally by Monte Carlo.

$\hat{\rho}_T$ is a super consistent estimator

It follows that $\hat{\rho}_T$ is a super consistent estimate of the true value.

$$\sqrt{T}(\hat{\rho}_T - 1) = \frac{T^{-3/2} \sum_{t=1}^{T} y_{t-1}\varepsilon_t}{T^{-2} \sum_{t=1}^{T} y_{t-1}^2}$$

but the numerator converges to $T^{-\frac{1}{2}}(1/2)\sigma^2[W(1)^2 - 1]$. The chi-square function has finite variance then the variance of the numerator is of order $(1/T)$, meaning that the numerator converges in probability to zero, hence $\sqrt{T}(\hat{\rho}_T - 1) \xrightarrow{P} 0$

Cointegration

In economics we usually think that there exist long-run relationships between many variables of interest. For example, although consumption and income may each follow random walks, it seem reasonable to expect that there is a long run relationship between these variables, or in other words that in the long run these variables move together. The alternative scenario will be that Income increase relative to consumption with time. This seems implausible. Other series that appear to move together are: short term - long term interest rates, imports and exports, prices and wages, stock prices and dividends ,etc.

Then the aim behind cointegration is the detection and analysis of long run relationships amongst economic time series variables. Given that most economic time series, appear to be non-stationary, they often require differencing or detrending to be transformed to stationarity. A problem with differencing or detrending is that we may remove relevant long run information. The cointegration analysis provides a way of retaining both short-run and long-run information.

Another reason why we are concerned with cointegration is that sometimes is thought to be a pre-requisite for the validity of some economic theory. For example if short term and long term interest rates (assuming there are I(1)) are not cointegrated, then, the term structure of interest rates cannot hold.

Definitions

A linear combination of two I(1) variables, say Y and X can be either I(1) or I(0). If this combination is I(1) the variables are said to be not-cointegrated. If there are I(0) such that $Y + \alpha X \tilde{} I(0)$ then the variables are said to be cointegrated.

Example 1

Consider the following model:

$$x_t + \beta y_t = u_t \tag{1}$$

$$x_t + \alpha y_t = e_t \tag{2}$$

$$u_t = u_{t-1} + \varepsilon_{1t} \tag{3}$$

$$e_t = \rho e_{t-1} + \varepsilon_{2t} \quad \text{with } |\rho| < 1 \tag{4}$$

$(\varepsilon_{1t}, \varepsilon_{2t})'$ is distributed identically and independently as a bivariate normal with

$$E(\varepsilon_{1t}) = E(\varepsilon_{2t}) = 0 \tag{5}$$

$$var(\varepsilon_{1t}) = \sigma_{11},\ var(\varepsilon_{2t}) = \sigma_{22},\ cov(\varepsilon_{1t}\varepsilon_{2t}) = \sigma_{12} \tag{6}$$

Solving for x_t and y_t form the above system with $\alpha \neq \beta$ gives

$$\begin{aligned} x_t &= \alpha(\alpha-\beta)^{-1}u_t - \beta(\alpha-\beta)^{-1}e_t, \\ y_t &= -(\alpha-\beta)^{-1}u_t + (\alpha-\beta)^{-1}e_t. \end{aligned}$$

Then we can conclude that both x_t and y_t are integrated of order one i.e., $x_t \tilde{} I(1)$, $y_t \tilde{} I(1)$, since u_t is integrated of order one. Nonetheless $x_t + \alpha y_t$ is I(0) because e_t is stationary. In this example the cointegrating vector is (1,α) and $x + \alpha y$ is the equilibrium relationship. In the long run the variables move towards the equilibrium $x + \alpha y = 0$ recognizing that this relationship need not to be realized exactly even as t tends to infinity.

In the bivariate case if the equilibrium condition exists, is unique.

Proof.

Suppose that there exist two distinct co-integrating parameters α and γ such that $x + \alpha y$ and $x + \gamma y$ are both ~I(0). This implies that $(\alpha - \gamma)y_t$ is also I(0) because a linear combination of two I(0) variable is also I(0). But we know that for $\alpha \neq \gamma$, $(\alpha - \gamma)y_t$ ~I(1) therefore we have a contradiction unless $\alpha = \gamma$.

Example

Consider the following example where cointegration of Prices and Dividends is a necessary condition for markets efficiency in the Fama sense.

Let us assume that stock prices might be written as

$$P_t = \sum_{i=1}^{\infty} (1/(1+r))^i E(D_{t+i}|I_t) + \varepsilon_t$$

where we may assume that ε_t is an I(0) process that might be, a white noise if it represents, say a measurement error, or it might be an autoregressive process if we assume agents are risk adverse. Let also assume D_t follows a random walk which is a special case of an I(1) variable.

$$D_t = D_{t-1} + \nu_t.$$

Then, we may express stock prices as

$$P_t = (1/r)D_t + \varepsilon_t$$

Given that D_t are integrated of order one, stock prices also are integrated of order one, since the sum of an I(1) process and an I(0) process is I(1).

We can easily see that if the theory is valid, $(1, -(1/r))$ is going to be a cointegrating vector since

$$(1, -(1/r)) \begin{bmatrix} P_t \\ D_t \end{bmatrix} = Z_t = \varepsilon_t$$

Notice that we assumed that ε_t was I(0), therefore if the theory holds dividends and prices should be cointegrated.

Different Representations for a cointegrating relationship

Consider the model described in equations (1) - (6). Take $|\rho| < 1$, then whenever x_t and y_t are cointegrated, we can show, for the simple two variables model, that the system of two equations has different representations, namely, vector autoregressive in levels, error-correction and moving-average representations.

VAR Representation

Let us reproduce for expositional reasons equations (1), (2), (3) and (4)

$$x_t + \beta y_t = u_t \tag{1}$$

$$x_t + \alpha y_t = e_t \tag{2}$$

$$u_t = u_{t-1} + \varepsilon_{1t} \tag{3}$$

$$e_t = \rho e_{t-1} + \varepsilon_{2t} \quad \text{with } |\rho| < 1 \tag{4}$$

If we lag one period equations (1) and (2) and subtract the lagged value from each expression we get

$$\Delta x_t + \beta \Delta y_t = \Delta u_t \tag{7}$$

$$\Delta x_t + \alpha \Delta y_t = \Delta e_t \tag{8}$$

Notice that

$\Delta u_t = \varepsilon_{1t}$

$\Delta e_t = -(1-\rho)e_{t-1} + \varepsilon_{2t}$ and using equation (2)

$= -(1-\rho)(x_{t-1} + \alpha y_{t-1}) + \varepsilon_{2t}$

then (7) and (8) might be rewritten as

$$\begin{bmatrix} 1 & \beta \\ 1 & \alpha \end{bmatrix} \begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t} \\ -(1-\rho)(x_{t-1} + \alpha y_{t-1}) + \varepsilon_{2t} \end{bmatrix}$$

and inverting the matrix we have

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \frac{1}{\alpha - \beta} \begin{bmatrix} \alpha & -\beta \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ -(1-\rho)(x_{t-1} + \alpha y_{t-1}) + \varepsilon_{2t} \end{bmatrix}$$

or

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \frac{1}{\alpha - \beta} \begin{bmatrix} \alpha\varepsilon_{1t} + \beta(1-\rho)(x_{t-1} + \alpha y_{t-1}) - \beta\varepsilon_{2t} \\ -\varepsilon_{1t} - (1-\rho)(x_{t-1} + \alpha y_{t-1}) + \varepsilon_{2t} \end{bmatrix} \tag{9}$$

or

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \frac{1}{\alpha - \beta} \begin{bmatrix} \beta(1-\rho) & \beta(1-\rho)\alpha \\ -(1-\rho) & -(1-\rho)\alpha \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \zeta_{1t} \\ \zeta_{2t} \end{bmatrix}$$

where
$\zeta_{1t} = (\alpha - \beta)^{-1}(\alpha\varepsilon_{1t} - \beta\varepsilon_{2t})$
$\zeta_{2t} = (\alpha - \beta)^{-1}(\varepsilon_{2t} - \varepsilon_{1t})$

Notice that this VAR representation IS NOT a VAR in the first differences. We would see later on that a VAR in first differences is not a possible representation when the variables are I(0). Notice that the var can be written as

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \frac{1}{\alpha - \beta} \begin{bmatrix} \beta(1-\rho)+1 & \beta(1-\rho)\alpha \\ -(1-\rho) & -(1-\rho)\alpha + 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \zeta_{1t} \\ \zeta_{2t} \end{bmatrix}$$

Error Correction Mechanism

In the 70's the way to proceed when the series under consideration where I(1) was just to take differences of these series and then regress the differenced series. It has been shown that this procedure is unsatisfactory mainly because it looses all long run information. (Also the interpretation of the coefficients is different from that one of the original regression)

In recent years one of the most popular ways of proceeding is to write our models as ECM (error correction mechanisms). The strong motivation for this models was mainly empirical since these models perform very well. In later years it have been shown that if two variables are cointegrated, there exist an ECM representation for these variables. This representation has the advantage that it keeps long run and short run information.

Consider a vector of $I(1)$ variables X_t, then if $X_t \tilde{} CI(1,1)$ there exist an error-correction representation for the data. A very general way of writing these type of models is

$$\phi(L)(1-L)X_t = -\alpha' X_{t-1} + \theta(L)\varepsilon_t,$$

where α is the cointegrating vector, $\theta(L)$ is a polynomial in the lag operator, $\phi(L)$ is a finite order lag polynomial with roots outside the unit circle and ε_t is a white noise. Also $\phi(0) = I$.

Alternatively we could write the model as

$$\Delta X_t = -\alpha' X_{t-1} + \Phi(L)\Delta X_t + \theta(L)\varepsilon_t,$$

where $\phi(L) = I + \Phi(L)$.

The intuition behind the error-correction model is that long run errors have to be corrected in the short run dynamics such that the process can move closer to its long run target.

Again let $X_t = (X_{1t}, X_{2t})$ and let Z_{t-1} be the cointegrating relationship.
Then the error correction equation for X_{2t} is

$$\Delta X_{2t} = \gamma Z_{t-1} + \Phi_1(L)\Delta X_{1t-1} + \Phi_2(L)\Delta X_{2t-1} + \theta(L)\nu_t$$

Then if the error correction theory is true we will expect that $\gamma < 0$ which implies that whenever $Z_{t-1} > 0$, *i.e.*, X_{2t} is above the long run equilibrium, then, ΔX_{2t} will be negative, *i.e.*, it will move in the direction of the equilibrium.

An important implication of the theory of cointegration is that asset prices cannot be cointerated if the weak efficiency market hypothesis holds. This is simply because if two asset prices are cointegrated, it is possible to forecast time t returns using information available a t time $t-1$. This goes against the simple no-arbitrage model.

ECM representation.

Consider once more the two variables model. It can be seen that the ECM representation follows directly from the VAR representation (equation (9)) .

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \frac{1}{\alpha - \beta} \begin{bmatrix} \beta(1-\rho)(e_{t-1}) \\ -(1-\rho)(e_{t-1}) \end{bmatrix} + \begin{bmatrix} \zeta_{1t} \\ \zeta_{2t} \end{bmatrix} \tag{9'}$$

Moving Average Representation
This follows directly from the VAR representation (1) and (2).
Equations (1) and (2) may be written in matrix notation as

$$\begin{bmatrix} 1 & \beta \\ 1 & \alpha \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} u_t \\ e_t \end{bmatrix},$$

which can be also written as

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \frac{1}{\alpha - \beta} \begin{bmatrix} \alpha & -\beta \\ -1 & 1 \end{bmatrix} \begin{bmatrix} u_t \\ e_t \end{bmatrix},$$

or taking first diferences as

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \frac{1}{\alpha - \beta} \begin{bmatrix} \alpha & -\beta \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \Delta u_t \\ \Delta e_t \end{bmatrix},$$

and noting that

$$\begin{bmatrix} \Delta u_t \\ \Delta e_t \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t} \\ (1-L)(1-\rho L)^{-1}\varepsilon_{2t} \end{bmatrix},$$

we can obtain the MA representation as

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \frac{1}{\alpha - \beta} \begin{bmatrix} \alpha & -\beta \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ (1-L)(1-\rho L)^{-1}\varepsilon_{2t} \end{bmatrix}.$$

We have analyzed the possible representations assuming a cointegration relationship between two variables. The concept of cointegration can be easily extended to n-variables.

Cointegration between n variables

Definition

Consider a n×1 vector of stochastic variables $X_t = (X_{1t}, X_{2t}, .., X_{n_t})$. We say that the elements of the vector are cointegrated of order (d, b), which we denote $X_t \sim \text{CI}(d, b)$ if

i) each of the components of X_t are I(d).

ii) there exists (at least) a vector such that $Z_t = \alpha' X_t$ is I$(d-b)$ for $d \geq b > 0$.

Then, α is called the cointegrating vector

If $d = b = 0$, then $\alpha' X_t = 0$ defines a long-run equilibrium relationship.

Notice that α is not unique since for any non-zero scalar b, $b\ \alpha' X_t$ is also integrated of order zero. Then in speaking of the cointegrating vector, an arbitrary normalization must be made, such that the first element of is unity.

If $n > 2$, there may be $r \leq n-1$ linearly independent $(n \times 1)$ vectors $(\alpha_1, ..., \alpha_r)$ such that $A' X_t \sim \text{I}(0)$, where A is the $(n \times r)$ matrix $A = [\alpha_1, ..., \alpha_r]$, such that rank$(A)$=$r$, where r is the cointegrating rank.

The vectors $(\alpha_1, ..., \alpha_r)$ are not unique, since for any non-zero $(r \times 1)$ vector b, $b' A' X_t \sim \text{I}(0)$, so $b' A'$ could be described as a cointegrating matrix.

Granger's Representation Theorem

In the two variables case we have shown that when two variables are cointegrated they do have a VAR, an ECM and a MA representation. These were particular cases of the Granger's Representation Theorem which is valid for a N variables vector.

Consider an n - vector time series X_t which satisfies:

$$\Phi(L) X_t = c + u_t$$

where $\Phi(L) = I_n - \sum_{i=1}^{P} \Phi_i L^i$, and u_t is a white noise with positive definite covariance matrix. It is assumed that det$[\Phi(z)]$= 0 which implies $|z| \geq 1$. Suppose that there exist exactly r cointegrating relationships among the elements of X_t . Then:

(*i*) there exists an $(n \times r)$ matrix A, of rank $r < n$ such that $A' X_t \sim \text{I}(0)$.

(*ii*) ΔX_t has an MA representation given by $\Delta X_t = \mu + \Psi(L) u_t$ with

$$A' \Psi(1) = 0,$$

where

$$\Psi(L) = I_n + \sum_{i=1}^{\infty} \Psi_i L^i$$

To understand the meaning of the restriction $A' \Psi(1) = 0$ and its implications consider the MA

$$\Delta X_t = \mu + \Psi(L)u_t$$

Now re-write it as $X_t = X_{t-1} + \mu + \Psi(L)u_t$ and substitute backwards to obtain

$$X_t = X_0 + \mu t + \Psi(1)\sum_{i=1}^{t} u_i + \eta_t - \eta_0$$

where $\mu = E(\Delta X_t)$ and $\{\eta_t\}$ is a I(0) sequence: $\eta_t = \sum_{S=0}^{\infty} a_s u_{t-s}, \qquad a_s = -\sum_{i=1}^{\infty} \Psi_{s+i}$

Then, pre-multiplying X_t by the cointegrating matrix A', we get the cointegrating relationship. Therefore each term in the right hand side has to be I(0).

$$A'X_t = A'(X_0 - \eta_0) + A'\mu t + A'\Psi(1)\sum_{i=1}^{t} u_i + A'\eta_t$$

Then, it is easy to see that for $A'X_t$ ~I(0), it is necessary that $A'\Psi(1) = 0$. This is only a necessary condition. Stationarity of $A'X_t$ further requires that $A'\mu = 0$. If $\mu \neq 0$, then unless $A'\mu = 0$ is satisfied, the linear combination $A'X_t$ will grow deterministically at rate $A'\mu$.

Notice that $A'\Psi(1) = 0$ implies that $\Psi(1)$ is singular since A' is a matrix with r LI vectors. This means that the matrix operator $\Psi(L)$ is non invertible. Thus, a cointegrated system can never be represented by a finite-order VAR for ΔX_t.

(iii) $\Phi(1) = BA'$

Theorem 1

The $(n \times n)$ matrix $\Phi(1)$ has a reduced rank $r < n$, and there exists an $(n \times r)$ matrix B such that $\Phi(1) = BA'$.

Proof. Consider an n-vector time series X_t which satisfies:

$$\Phi(L)X_t = c + u_t$$

and the Wold representation

$$\Delta X_t = \mu + \Psi(L)u_t.$$

Then multiplying the Wold representation by $\Phi(L)$, we get

$$(1-L)\Phi(L)X_t = \Phi(1)\mu + \Phi(L)\Psi(L)u_t$$

and using the autoregressive representation (multiplied by$(1-L)$) we get

$$(1-L)\Phi(L)X_t = (1-L)(c + u_t)$$

and equating the last two expressions we get

$$(1-L)u_t = \Phi(1)\mu + \Phi(L)\Psi(L)u_t$$

(since $(1-L)c = 0$) then, the above expresion implies that:
1) $\Phi(1)\mu = 0$.
2) $(1-L) = \Phi(L)\Psi(L)$, in particular for $L = 1$, requires that $\Phi(1)\Psi(1) = 0$.

Let π' denote a row of $\Phi(1)$, then conditions 1) and 2) imply that π is a cointegrating vector.

Now if $a_1, a_2, ...a_r$ form a basis for the space of cointegrating vectors, then we can express, $\pi_{nx1} = (a_1, a_2, ...a_r)_{nxr} b_{r\times 1}$, since any linear combination of a cointegrating vector is also a cointegrating vector.or $\pi' = b'A'$ Applying the same reasoning for each of the rows of the rows we get

$$\Phi(1) = BA'$$

■

(*iv*) There exist a VAR representation in Levels and the determinant of the polynomial in Φ has a unit root. This can be shown by noticing that $\Phi(1)$ is singular.

(*v*) *Error Correction Representation.*

To show this point we first need to transform the original VAR

$$\Phi(L)X_t = c + u_t,$$

using the following relationship

$$I_n - \sum_{i=1}^{p} \Phi_i L^i = I_n - (\sum_{i=1}^{P} \Phi_i)L - (I-L)\sum_{i=1}^{p-1} \Gamma_i L^i$$

where
$\Gamma_i = -\sum_{i=j}^{p-1} \Phi_{j+1}$ for $j = 1, ..., p-1$
Then

$$(I_n - \sum_{i=1}^{p} \Phi_i L^i)X_t = \left(I_n - \left(\sum_{i=1}^{p} \Phi_i\right)L + (I-L)\sum_{i=1}^{p-1} \Gamma_i L^i\right)\right) X_t = c + u_t.$$

Then rearranging terms

$$\begin{aligned} X_t &= c + (\sum_{i=1}^{p} \Phi_i)X_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + u_t \\ &= c + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + (I_n - \Phi(1))\, X_{t-1} + u_t \end{aligned}$$

$$\Delta X_t = c + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} - \Phi(1) X_{t-1} + u_t$$

From this expression we can derive the following results.

(a) If rank $[\Phi(1)] = 0$ then $\Phi(1) = 0$ and $X_t \tilde{} I(1)$ since we can write everything in terms of a VAR in diferences. This is only valid when the variables do not cointegrate.

(b) If rank $[\Phi(1)] = n$ then $det(\Phi(1)) \neq 0$ ($\Phi(L)$ does not have a unit root). This implies that $X_t \tilde{} I(0)$ and therefore it should have a MA representation since we can invert the original expression, *i.e.*,

$$X_t = \Phi(L)^{-1}(c + + u_t)$$

(c) If rank $[\Phi(1)] = r$, $0 < r < n$ then $\Phi(1) = BA'$, where B is an nxr matrix .

(See a explanation of these in the section about Johansen Procedure).

Restrictions in the parameters of the Error Correction Representation.

$$\Delta X_t = c + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} - B Z_{t-1} + u_t, \; Z_t = A' X_t$$

In this error-correction representation all variables are $I(0)$. Then the term $A'X_t$ is viewed as the "error" from the long run equilibrium relationship, and B gives the "correction" to X_t caused by this error.

Notice that taking expected values in both sides we get;

$$[I - \sum_{i=1}^{p-1} \Gamma_i L^i] E(\Delta X_t) = c - B E(Z_{t-1})$$

Thus, in order to have a system in which there is no drift in any of the variables [i.e., $E(\Delta X_t) = 0$], we need to impose the restriction $c = BE(Z_{t-1})$ (this is equivalent to $A'\mu = 0$) In the absence of such a restriction, it is implied by the ECM that there are $n - r$ separate time trends that account for the trend X_t .

Then imposing this restriction we obtain

$$\begin{aligned} \Delta X_t &= BE(Z_{t-1}) + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + -B Z_{t-1} + u_t \\ &= -B(Z_{t-1} - E(Z_{t-1})) + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + u_t \end{aligned}$$

the intercept enters the system only via the error-correction term and there is no autonomous growth component.

Again notice that A is not unique since $\Phi(1) = BA' = BPP^{-1}A' = B^*A^{*\prime}$, for all $r \times r$ non-singular matrices P.

Tests For Cointegration.

Univariate Tests for Cointegration.

1) Two Stages Approaches

First stage: A static Regression
For the model

$$y_t = \beta x_t + \varepsilon_t$$

Regress y on x using OLS achieves a consistent estimate of the long-run steady -state relationship between the variables of the model, and all dynamics and endogeneity issues can be ignored asymptotically. This arises because of the super consistency property of the OLS estimator when the series are cointegrated.

Suppose that the model that dynamic model that captures both short adjustment and the long run relationship is

$$y_t = \gamma_o x_t + \gamma_1 x_{t-1} + \alpha y_{t-1} + \varepsilon_t$$

This can be re-written as

$$y_t = \lambda_0 x_t + \lambda_1 \Delta x_t + \lambda_2 \Delta y_t + \varepsilon_t$$

where $\lambda_0 = \frac{\gamma_0+\gamma_1}{1-\alpha}$, $\lambda_1 = \frac{-\gamma_1}{1-\alpha}$, $\lambda_2 = \frac{-\alpha}{1-\alpha}$.

Thus, estimating the static model to obtain the long-run parameter β is equivalent to estimating the dynamic model without the short run terms. According to the super consistency property if y_t and x_t are both non-stationary I(1) variables, and ε_t ~I(0), then as the sample size, T, becomes larger the OLS estimator converges to its true value at a much faster rate than the I(0) variables. Of course the omitted dynamic terms are captured by the residuals. This we will see later will be a problem in short samples.

As a second stage I can use alternative testing stragesies:

i) **Make an (ADF) test for unit roots for the residuals:***: The Engle - Granger Approach.*

If you do not reject the Hypothesis that the residuals have a unit root (that there are I(1)), then Y and X are not cointegrated. On the other hand if you do reject this hypothesis against the stationary alternative, then you conclude that the residuals (a linear combination of Y and X) are I(0), then Y and X are cointegrated.

Therefore we regress

$$\Delta\hat{\varepsilon}_t = \phi\hat{\varepsilon}_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta\hat{\varepsilon}_{t-i} + \mu + \delta t + \zeta_t$$

The question of including a trend and or a constant term in the test regression depends of whether this terms appear in the original regression. That is the deterministic component can be added either to the static regression or to the ADF regression, but not to both. Hansen (1992) has shown that including a deterministic trend results in loss of power.[1]

Note that is not possible to use standard Dickey - Fuller critical values. The standard table will tend to over reject the null of no cointegration. Note also that the distribution of the test statistic under the null is affected by the number of regressors included in the static model.

Fortunately MacKinon (1991) provides Table which has linked the critical values for particular tests to a set of parameters of an equation of response surfaces.

$$C(p) = \phi_\infty + \phi_1 T^{-1} + \phi_2 T^{-2}$$

Response surfaces for critical values of cointegration tests.

[1] Notice that if we consider the simple DF test for the residuals

$$\Delta\hat{\varepsilon}_t = \phi\hat{\varepsilon}_{t-1} + \zeta_t$$

This can be re-written (evaluating at $\beta = \hat{\beta}$)

$$\Delta(y_t - \beta x_t) = \phi(y_{t-1} - \beta x_{t-1}) + \zeta_t$$

or

$$\Delta y_t = \beta\Delta x_t + \phi(y_{t-1} - \beta x_{t-1}) + \zeta_t.$$

This model is not an unrestricted ECM and it imposes that the change in the short run is the same that the long run effect. This is unlikely to be true.

n	Model	%point	ϕ_∞	ϕ_1	ϕ_2
1	no C	1	-2.5658	-1.960	-10.04
	no t	5	-1.9393	-0.398	0.0
		10	-1.6156	-0.181	0.0
1	C	1	-3.4336	-5.999	-29.25
	no t	5	-2.8621	-2.738	-8.36
		10	-2.5671	-1.438	-4.48
1	C	1	-3.9638	-8.353	-47.44
	t	5	-3.4126	-4.039	-17.83
		10	-3.1279	-2.418	-7.58
3	C	1	-4.2981	-13.790	-46.37
	no t	5	-3.7429	-8.352	-13.41
		10	-3.4556	-6.241	-2.79

where $C(p)$ is the p per cent critical value.

The Cointegrating Durbin-Watson.

ii) **The null may be tested using the Sargan-Bhargava or CDW test**

This test is very simple and consist in comparing the DW statistic with tabulated values. The rationality of this procedure is as follows: Consider equations (2) and (4).

$$x_t + \alpha y_t = e_t \tag{2}$$

$$e_t = \rho e_{t-1} + \varepsilon_{2t} \quad \text{with } |\rho| < 1 \tag{4}$$

A regression of x_t on y_t will yield serially correlated residuals. We can use the DW statistic to get information about ρ since this statistic is approximately $2(1-\rho)$.

$$DW \cong 2(1-\rho)$$

Then if ρ is equal to 1 (a unit root), the DW statistic equals zero. Sargan and Bhargava provide tables to test this hypothesis. However, this critical value is only relevant when the disturbance follows a first order autoregressive process and there is there no higher serial correlation, which is unlikely. Thus the CRDW test is generally not a suitable test statistic.

2) Three Stages:The Engle-Granger-Yoo Approach.

Engle and Yoo propose a third step to the standard Engle-Granger procedure which seeks to overcome some of the problems inherent in using the static model which yields β generally biased in small samples. Assuming that there is a unique cointegration vector and weak exogeneity of the short run parameters, then the third step provides a correction of the first stage estimate of β and ensures it has a normal distribution. The methodology consists of correcting the long-run relationship by the small sample bias $(\gamma/(1-\alpha))$ and use the correct residuals to perform the ADF test.

Stg1 Regress $y_t = \gamma_o x_t + \gamma_1 x_{t-1} + \alpha y_{t-1} + \varepsilon_t$

Stg2 Construct $\widehat{u}_t = y_t - \frac{\widehat{\gamma}_0 + \widehat{\gamma}_1}{1-\widehat{\alpha}} x_t$

Stg3 Perform an ADF test for $\widehat{u}_t$.

Cointegration in Multivariate Systems- The Johansen Approach

Defining z_t , a vector of n potentially endogenous variables it is possible to specify the following DGP and the model z_t as an unrestricted VAR involving k-lags of z_t

$$z_t = A_1 z_{t-1} + ... + A_k z_{t-k} + u_t$$

where z_t is $n \times 1$ and each of the A_i is $n \times n$ matrix of parameters. As we show above this equation can be written as a vector error-correction model

$$\Delta z_t = \Gamma_1 \Delta z_{t-1} + ... + \Gamma_{k-1} \Delta z_{t-k-1} + \Pi z_{t-k} + u_t$$

where $\Gamma_i = -(I - A_1 - ... - A_i)$ $(i = 1, .., k-1)$, and $\Pi = -(I - A_1 - ... - A_k)$.[2] This way of specifying the system contains information on both the short run and the long run adjustments to changes in z_t , via the estimates of $\hat{\Gamma}_i$ and $\hat{\Pi}$ respectively. As it will be seen, $\Pi = \alpha\beta'$, where α represents the speed of adjustment to disequilibrium, while β is a matrix of long-run coefficients such that the term $\beta' z_{t-k}$ represents up to $n-1$ cointegration relationships. Assuming that z_t is a vector of non-stationary I(1) variables, then all the terms in differences are I(0) while Πz_{t-k} must also be I(0) for the error term to be a white noise.

There are two cases where this requirement is met:

1. When there is no cointegration at all, implying that there are no linear combinations of z_t that are I(0), and consequently Π is a matrix of $n \times n$ zeros. In this case the appropriate model is a VAR in differences.

[2] Notice that in this presentation the variable in levels is LAG k, while in the previous presentation the variable in levels was lag 1. The analysis can be carried out in both ways.

2. When there exist up to $n-1$ cointegration relationships $\beta' z_{t-k}$ ~I(0). In this instance there will exist a number $r \leq n-1$ cointegration vectors where r is the number of columns of which form r LI combinations of the variables in z_t , together with $(n-r)$ non-stationary vectors in β. Only the cointegration vectors in enter in the VECM, which implies that for the last $(n-r)$ columns, the are insignificantly small. Thus the typical problem faced, of determining how many r cointegration vectors exist amounts to equivalently testing which columns in Π are zero. Consequently testing for cointegration amounts to a consideration of the rank of Π, that is, finding the number of r linearly independent columns in Π.

$-If\ \Pi$ *is full rank the variables in* z_t *have to be* $I(0)$

$-If\ \Pi$ *is zero there is no cointegrating vector*

$-If\ \Pi$ *is reduced rank the number of cointegrating vectors is the* $RANK(\Pi)$

Canonical Correlations

Population Canonical Correlations

Let the $(n_1 \times 1)$ vector y_t and the $(n_2 \times 1)$ vector x_t denote stationary random variables. Typically this variables are measured in deviations from the population mean such that $E(y_t y_t')$ represents the variance-covariance matrix of y_t .

In general

$$\begin{bmatrix} E(y_t y_t') & E(y_t x_t') \\ E(x_t y_t') & E(x_t x_t') \end{bmatrix} = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix}$$

We can often gain some insight into the nature of these correlations by defining two new $(n \times 1)$ random vectors φ_t and ξ_t , where n is the smaller of n_1 and n_2 .

$$\begin{aligned} \varphi_t &= K' y_t \\ \xi_t &= A' x_t \end{aligned}$$

The matrices K' and A' are chosen such that

$$\begin{bmatrix} E(y_t y_t') & E(y_t x_t') \\ E(x_t y_t') & E(x_t x_t') \end{bmatrix} = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix}$$

$E(\varphi_t \varphi_t') = K' \Sigma_{YY} K = I,$

$E(\xi_t \xi_t') = A' \Sigma_{XX} A = I$ and,

$$E(\varphi_t \xi_t') = R = \begin{bmatrix} r_1 & & 0 \\ & r_2 & \\ & & \\ 0 & & r_n \end{bmatrix}$$

where the elements of φ_t and ξ_t are ordered in such a way that$1 \geq r_1 \geq r_2..... \geq r_n \geq 0$

The population r_i is known as the i^{th} population canonical correlation between y_t and x_t .

The canonical correlations can be calculated by calculating the eigen values of

$$\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY},$$

$\lambda_1 > \lambda_2 > ...\lambda_n$, and the canonical correlations turn out to be the square roots of these eigenvalues.

The Johansen Method of reduced rank regression

The procedure may be described as follows:

First rewrite the ECM Equation

$$\Delta z_t + \alpha\beta' z_{t-k} \quad = \Gamma_1 \Delta z_{t-1} + ... + \Gamma_{k-1}\Delta z_{t-(k-1)} + u_t$$

it is possible to correct for the short run dynamics (i.e., take out their effect) by regressing Δz_t and $'z_{t-k}$ separately on the right hand side of the previous equation, i.e.,

$$\begin{aligned} \Delta z_t &= P_1 \Delta z_{t-1} + ... + P_{k-1}\Delta z_{t-(k-1)} + R_{0t} \\ z_{t-k} &= T_1 \Delta z_{t-1} + ... + T_{k-1}\Delta z_{t-(k-1)} + R_{kt} \end{aligned}$$

Which can be used to form the residual (product moment) matrices

$$\widehat{S}_{ij} = T^{-1}\sum \widehat{R}_{it}\widehat{R}'_{jt} \quad i,j = 0,k$$

The maximum likelihood estimate of β is obtained as eigenvectors corresponding to the largest eigenvalues from solving the equation

$$|\lambda \widehat{S}_{kk} - \widehat{S}_{0k}\widehat{S}_{00}^{-1}\widehat{S}_{0k}| = 0$$

which gives the n eigenvalues $\hat{\lambda}_1 > \hat{\lambda}_2 > .....\hat{\lambda}_n$ and their corresponding eigen vectors, $\hat{\varphi} = (\hat{\varphi}_1 > \hat{\varphi}_2 > ... > \hat{\varphi}_n)$. Those r elements in $\hat{\varphi}$ which determine linear combinations of stationary relationships can be denoted $\beta = (\hat{\varphi}_1, \hat{\varphi}_2, ..., \hat{\varphi}_r)$, that is, these are the cointegration vectors. This is because the eigen values are the largest squared canonical correlations between the "levels" residuals R_{kt} and the "differences" residuals R_{0t}, that is, we obtain estimates of all the distinct

$\hat{\varphi}_i' z_t$ combinations of the I(1) levels of z_t which produce high correlations with the stationary Δz_t elements, such combinations being the cointegrating vectors by the virtue of the fact that they must themselves be I(0) to achieve a high correlation. Thus the magnitude of $\hat{\lambda}_i$ is a measure of how strongly the cointegration relations are correlated with the stationary part of the model. The last $(n-r)$ combinations indicate the non-stationary combinations and theoretically are uncorrelated with the stationary elements. Consequently, for the eigenvectors corresponding to the non-stationary part of the model, $\hat{\lambda}_i = 0$ for $i = r+1, ..., n$.

Testing for reduced rank

To find the number of cointegrating vectors we said that is equivalent to find the number of linearly independent columns in Π or the number of $n-r$ columns of significantly small.

The approach amounts to a reduced rank regression which provides n eigenvalues $\hat{\lambda}_1 > \hat{\lambda}_2 > .....\hat{\lambda}_n$ and their corresponding eigen vectors,$\hat{\varphi} = (\hat{\varphi}_1 > \hat{\varphi}_2 > ... > \hat{\varphi}_n)$. Those r elements in $\hat{\varphi}$ which determine linear combinations of stationary relationships can be denoted $\beta = (\hat{\varphi}_1, \hat{\varphi}_2, ..., \hat{\varphi}_r)$ that is, the distinct $\hat{\varphi}_i' z_t$, which we will denote $\beta_i' z_t$, are correlated with the stationary part of the model. The last $n-r$ combinations obtained from the Johansen approach indicate the non stationary combinations, and theoretically these are uncorrelated with the stationary elements in the ECM. Consequently, for the eigenvectors corresponding to the non-stationary part of the model, $\hat{\lambda}_i = 0$ for $i = r+1, ..., n$.

Thus to test the null hypothesis that there are at most r cointegration vectors amounts to test

$$H_0)\lambda_i = 0 \quad i = r+1, ..., n.$$

where only the first r eigenvalues are non-zero. (This is tested against the alternative of n cointegrating vectors).

It can be shown (see Hamilton) that the likelihood test that corresponds to the ECM under the null that there are only r cointegrating vectors is

$$L^*(H_0) = -(Tn/2)log(2\pi) - (Tn/2) - (T/2)log\left|\widehat{S}_{00}\right| - (T/2)\sum_{i=1}^{r}\log(1-\hat{\lambda}_i).$$

It can also be shown that the likelihood test that corresponds to the ECM without any restriction on the number of cointegrating vectors is

$$L^* = -(Tn/2)log(2\pi) - (Tn/2) - (T/2)log\left|\widehat{S}_{00}\right| - (T/2)\sum_{i=1}^{n}\log(1-\hat{\lambda}_i).$$

Then a likelihood ratio test, *using a non standard distribution*, can be constructed, using what is known as the **Trace statistic.**

$$\lambda_{trace} = -T \sum_{i=r+1}^{n} \log(1 - \hat{\lambda}_i) \qquad r = 0, 1, ..., n-2, n-1$$

Another test of the significance of the largest λ_r is the so called maximal-eigenvalue or $\boldsymbol{\lambda - \max}$ **statistic** :

$$\lambda_{\max} = -T \log(1 - \hat{\lambda}_{r+1}) \quad r = 0, 1, ..., n-2, n-1.$$

This tests the existence of r cointegrating vectors against the alternative that $r+1$ exist and is derived in exactly same way.

Testing restrictions on the cointegrating vector.

Many times we are interested to test restrictions on the cointegrating vector. For example I might be interested in some theoretical long run relationship which impose some restrictions on the values of the cointegrating relationship. We may be also interested in testing whether we should include or not a regressor in the cointegrating relationship.

The crucial point in deriving a LR test for these type of hypothesis is that both under the null and the alternative there are r cointegrating relationships and therefore the asymptotic theory is standard since the regressions only involve variables which are I(0)and the test would be distributed *chi-square*. The LR test will be

$$LR = -T \sum_{i=1}^{r} \log(1 - \hat{\lambda}_i) + T \sum_{i=1}^{r} \log(1 - \tilde{\lambda}_i)$$

Arch Models

Most investors dislike risk taking and require a premium for holding assets with risky payoffs. The variance of an asset has been used to measure risk, and split the risk into a company specific component, which is diversifiable, and a market component which cannot be diversified. This measure of the unconditional volatility does not recognize that there may be predictable patterns in stock market volatility. We will analyze models of conditional (on information at time t-1) volatility. These type of models have the implication for finance that investors can predict the risk. This type of models successfully characterize the fact that stock prices seem to go through long periods of high and long periods of low volatility.

The fact that market participants may predict volatility has important implications The most important is that for periods where the investor has forecasted prices to be very volatile, she should either exit the market or require a large premium as a compensation for bearing an unusual high risk.

Empirical Regularities of Asset Returns.

i Thick Tails

Asset returns tend to be leptokurtotic . The documentation of this empirical regularity is presented in Mandelbrot (1965).

ii Volatility Clustering

" ... large changes tend to be followed by large changes, of either sign and small changes tend to be followed by small changes "

iii Leverage Effects

The so-called "leverage effect" first noted by Black(1976) refers to the tendency for stock prices to be negatively correlated with changes in stock volatility. A firm with debt and equity outstanding typically becomes more highly leveraged when the value of the firm falls. This raises the equity return volatility.

iv) Non-Trading Periods

Information that accumulates when financial markets are closed is reflected in prices after the markets reopen. If for example, information accumulates at a constant rate over calendar time, then the variance of the returns over the period from Friday close to the Monday close should be three times the variance from the Monday close to the Tuesday close.

v) Forecastable Events

Patell and Wolfson (1979,1981) show that individual firm's stock returns volatility is high around earning announcements.

Introduction: Conditional and Unconditional moments

Before presenting the alternative Arch type models, we will briefly review the difference between conditional and unconditional moments.

Let us assume that y_t follows a random walk, i.e.

$$y_t = y_{t-1} + \varepsilon_t$$

Then

$$y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i$$

Unconditional Moments

The unconditional mean and variance are;

$$E(y_t) = y_0$$
$$V(y_t) = t\sigma^2$$

A RW has a constant unconditional mean but a time varying unconditional variance.

Conditional Moments

The conditional mean and variance are;

$$E(y_t|y_{t-1}) = y_{t-1}$$
$$V(y_t|y_{t-1}) = E(y_t - E(y_t|y_{t-1}))^2 = E(y_{t-1} + \varepsilon_t - E(y_t|y_{t-1}))^2 = \sigma^2$$

A RW has a constant unconditional mean but a time varying unconditional variance.

So while the unconditional variance of a random walk model tends to infinite as t increase, the conditional variance is constant.

Univariate Parametric Models

Arch Models

In the linear Arch(q) model originally introduced by Engle(1982), the time varying conditional variance is postulated to be a linear function of the past q squared innovations.

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 = w + \alpha(L)\varepsilon_{t-1}^2$$

A sufficient condition for the conditional variance to be positive is that the parameters of the model satisfy the following constraints; $\omega > 0$ and $\alpha_1 > 0, \alpha_2 > 0, , ...\alpha_q > 0$

Defining $\nu_t \equiv \varepsilon_t^2 - \sigma_t^2$, the ARCH(q) model can be re-written as

$$\varepsilon_t^2 = \omega + \alpha(L)\varepsilon_{t-1}^2 + \nu_t$$

(Notice that $\sigma_t^2 = E(\varepsilon_t^2 | \varepsilon_{t-1}^2, \varepsilon_{t-2}^2, ...)$) Since $E_{t-1}(\nu_t) = 0$, the model corresponds to an AR(q) model for the squared innovations, ε_t^2 . Then, the process is covariance stationary if and only if the sum of the positive autoregressive parameters is less than one, in which case the unconditional variance equals

$$Var(\varepsilon_t^2) = \omega/(1 - \alpha_1 - \alpha_2 ... - \alpha_q).$$

Even though ε_t 's are serially uncorrelated they are clearly not independent through time. In accordance with the stylized facts for assets returns discussed above, there is a tendency for large (small) absolute values of the process to be followed by other large (small) values of unpredictable sign.

The ARCH(1) Model

Constant unconditional Variance but non-constant conditional Variance.

Some useful statistical results are given below for the simplest ARCH(1) model, which is identical to the one used by Engle (1982). The main result is that this simple model exhibits constant unconditional variance but non-constant conditional variance.

Consider the following model

$$y_t = \mu + \varepsilon_t$$

$$\varepsilon_t = u_t(\omega + \alpha\varepsilon_{t-1}^2)^{1/2}, u_t \tilde{} IIN(0,1), \omega > 0, \alpha > 0$$

(NOTICE that $(\omega + \alpha\varepsilon_{t-1}^2)^{1/2}$ is the conditional standard deviation, σ_t defined as $(E(\varepsilon_t^2 | \varepsilon_{t-1}^2, \varepsilon_{t-2}^2, ...))^{1/2}$.

i) The conditional expectation of ε_t is equal to zero

$$E(\varepsilon_t | \varepsilon_{t-1}) = E(u_t | \varepsilon_{t-1})(\omega + \alpha\varepsilon_{t-1}^2)^{1/2} = 0$$

Notice that $E(u_t|\varepsilon_{t-1}) = E(u_t) = 0$, since u_t ~IIN(0,1)

ii) The conditional variance is given by the following formula

$$Var(\varepsilon_t|\varepsilon_{t-1}) = E(u_t^2|\varepsilon_{t-1})(\omega + \alpha\varepsilon_{t-1}^2) = (\omega + \alpha\varepsilon_{t-1}^2)$$

Notice that $E(u_t^2|\varepsilon_{t-1}) = E(u_t^2) = 1$, since u_t ~IIN(0,1)

Then the conditional mean and variance of y_t are given by the following formulae;

$$E(y_t|y_{t-1}) = \mu$$

$$Var(y_t|y_{t-1}) = (\omega + \alpha\varepsilon_{t-1}^2)$$

Then, the conditional variance of y_t is time varying. On the other hand it can be easily seen that the unconditional variance is time invariant whenever ε_t^2 is stationary, i.e.

$$V(y_t) = V(\varepsilon_t) = \omega/(1-\alpha)$$

whenever the process is stationary.

(since $V(\varepsilon_t) = E(\varepsilon_t^2) = E(\omega + \alpha\varepsilon_{t-1}^2) = \omega + \alpha E(\varepsilon_{t-1}^2)$)

First Order Autoregressive Process with ARCH effects.

For more complicated models such as AR(1)-ARCH(1), we obtain similar results provided that the process for y is stationary, i.e. that the autoregressive parameter is smaller than one in absolute value.

Assume the following first order autoregressive process

$$y_t = \theta y_{t-1} + \varepsilon_t$$

where $\varepsilon_t = u_t(\omega + \alpha\varepsilon_{t-1}^2)$ and u_t ~IIN(0,1) , $\omega > 0$, $\alpha > 0$

then

i) The conditional expectation of ε_t is equal to zero

$$E(\varepsilon_t|\varepsilon_{t-1}) = E(u_t^2|\varepsilon_{t-1})(\omega + \alpha\varepsilon_{t-1}^2) = 0 \; since \; E(u_t|\varepsilon_{t-1}) = E(u_t) = 0$$

ii) The conditional variance is given by the following formula

$$Var(\varepsilon_t|\varepsilon_{t-1}) = E(u_t^2|\varepsilon_{t-1})(\omega + \alpha\varepsilon_{t-1}^2) = (\omega + \alpha\varepsilon_{t-1}^2)$$

since $E(u_t^2|\varepsilon_{t-1}) = E(u_t) = 1$

Then the conditional mean and variance of y_t are given by the following formulae;

$$E(y_t|y_{t-1}) = \theta y_{t-1}$$

$$Var(y_t|y_{t-1}) = (\omega + \alpha\varepsilon_{t-1}^2)$$

To find the unconditional variance of y_t we recall the following property for the variance;

$$Var(y_t) = E(Var(y_t|y_{t-1})) + Var(E(y_t|y_{t-1}))$$

then

i) $E(Var(y_t|y_{t-1})) = E(\omega + \alpha\varepsilon_{t-1}^2) = \omega + \alpha E(\varepsilon_{t-1}^2) = \omega + \alpha Var(\varepsilon_{t-1})$

ii) $Var(E(y_t|y_{t-1})) = \theta^2 Var(y_{t-1})$

Then if the process is covariance stationary we have

$$\begin{aligned} Var(y_t) &= \frac{\omega + \alpha Var(\varepsilon_{t-1})}{(1-\theta^2)} \\ &= \frac{\omega}{(1-\alpha)(1-\theta^2)} \end{aligned}$$

(Since $Var(\varepsilon_{t-1}) = \omega/((1-\alpha))$

GARCH Models

In empirical applications it is often difficult to estimate models with large number of parameters, say ARCH(q). To circumvent this problem Bollerslev (1986) proposed the Generalized ARCH or GARCH(p, q) model,

$$\begin{aligned} \sigma_t^2 &= \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \\ &= \omega + \alpha(L)\varepsilon_{t-1}^2 + \beta(L)\sigma_{t-1}^2 \end{aligned}$$

A sufficient condition for the conditional variance in the GARCH(p, q) model to be well defined is that all the coefficients in the infinite order linear ARCH model must be positive. Provided that $\alpha(L)$ and $\beta(L)$ have no common roots and that the roots of the polynomial in L, $(1 - \beta(L)) = 0$ lie outside the unit circle, this positive constraint is satisfied, if and only if, the coefficients of the infinite power series expansion for $\alpha(L)/(1 - \beta(L))$ are non-negative.

Rearranging the GARCH(p, q) model by defining $\nu_t \equiv \varepsilon_t^2 - \sigma_t^2$, it follows that

$$\varepsilon_t^2 = \omega + (\alpha(L) + \beta(L))\varepsilon_{t-1}^2 - \beta(L)\nu_{t-1} + \nu_t$$

which defines an ARMA(Max(p, q),p) model for ε_t^2

By standard arguments, the model is covariance stationary if and only if all the roots of $(1 - \alpha(L) - \beta(L))$ lie outside the unit circle.

If all the coefficients are positive, this is equivalent to the sum or the autoregressive coefficients being smaller that 1.

The analogy to ARMA class of models also allows for the use of standard time series techniques in the identification of the orders of p and q.

In most empirical applications with finitely sampled data, the simple GARCH$(1, 1)$ is found to provide a fair description of the data.

Persistence and Stationarity

Using the GARCH(1,1) model it is easy to construct multi period forecasts of volatility. When $\alpha+\beta < 1$, the unconditional variance of ε_{t+1} is $\omega/(1-\alpha-\beta)$.

If we re-write the following GARCH(1,1) as

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \\ &= \omega + \alpha(\varepsilon_{t-1}^2 - \sigma_{t-1}^2) + (\alpha + \beta)\sigma_{t-1}^2 \end{aligned}$$

The coefficient measures the extent to which a volatility shock today feeds through into next periods volatility, while $(\alpha + \beta)$ measures the rate at which this effect dies over time. Recursively substituting and using the law of iterated expectations, the conditional expectation of volatility j periods ahead is,

$$E_t[\sigma_{t+j}^2] = (\alpha + \beta)^j(\sigma_t^2 - \omega/(1 - \alpha - \beta)) + w/(1 - \alpha - \beta)$$

The multi period volatility forecast reverts to its unconditional mean at rate $(\alpha + \beta)$.

IGARCH Models

Integrated GARCH models are processes were the autorregresive part of the square residuals has a unit root, i.e., $(\alpha + \beta) = 1$. For this case the conditional expectation of the volatility j periods ahead is

$$E_t[\sigma_{t+j}^2] = \sigma_t^2 + j\omega.$$

This process looks very much as a random walk with drift ω. Then, if ε_t follows an IGARCH process the unconditional variance does not exist and therefore it is not covariance stationary. Nelson(1990) shows that the analogy with the random walk process should be treated with caution since the IGARCH process is not covariance stationary but it may be proved to be strictly stationary.

For example when $\omega = 0$, $E_t[\sigma^2_{t+j}] = \sigma^2_t$, so volatility is a martingale. But the volatility remains bounded , since it cannot be negative, and then using the fact that a bounded martingale must converge, we can show that it converges to zero, a degenerate distribution.

Despite the fact that it seems to be an empirical regularity that volatility is IGARCH (many estimated models have coefficients that sum near 1) we regard this type of process as unlikely. (see section on structural breaks and GARCH models)

EGARCH Models

Even if GARCH models successfully capture thick tailed returns, and volatility clustering, are not well suited to capture the "leverage effect" since the conditional variance is a function only of the magnitudes of the lagged residuals and not their signs

In the exponential GARCH (EGARCH) model of Nelson (1991) σ^2_t depends on both the size and the sign of lagged residuals.

EGARCH(1,1) Models

$$\ln \sigma^2_t = \alpha_0 + \beta_1 \ln \sigma^2_{t-1} + \gamma_0([|\varepsilon_{t-1}/\sigma_{t-1}| - (2/\pi)^{1/2}] + \delta(\varepsilon_{t-1}/\sigma_{t-1}))$$

Obviously the EGARCH model always produces a positive conditional variance σ^2_t for any choice of α_0, β_1, γ_0 and so that no restrictions need to be placed on these coefficients (except $|\beta_1| < 1$). Because of the use of both $|\varepsilon_t/\sigma_t|$ and (ε_t/σ_t), σ^2_t will also be non-symmetric in ε_t and, for negative δ, will exhibit higher volatility for large negative ε_t .

Other ARCH Specifications

Glosten, Jagannathan and Runkle (1989) proposed the following specification:

$\varepsilon_t = \sigma_t \nu_t$, where v_t is iid.

$$\sigma^2_t = \alpha_0 + \beta_1 \sigma^2_{t-1} + \alpha_1 \varepsilon^2_{t-1} + \alpha_2 \varepsilon^2_{t-1} I_{t-1},$$

where, $I_{t-1} = 1$ if $\varepsilon_{t-1} \geq 0$ and $I_{t-1} = 0$ if $\varepsilon_{t-1} < 0$.

The non-negativity condition is satisfied provided that all the parameters are positive. If leverage effects do exist, $\alpha_2 < 0$.

Additional Explanatory Variables.

It is straightforward to add other explanatory variables to a GARCH specification. Glosten, Jagannathan and Runkle(1993) add a short-term nominal interest rate to various GARCH models and show that it has a significant positive effect on stock market volatility.

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2 + \gamma X_{t-1},$$

where X is any positive variable.

GARCH in Mean Models

Many theories in finance assume some kind of relationship between the mean of a return and its variance . A way to take this into account is to explicitly write the returns as a function of the conditional variance or, in other words, to include the conditional variance as another regressor. *GARCH in Mean Models* allow for the conditional variance to have mean effects. Most of the time this conditional variance term will have the interpretation of a time varying risk premium.

Consider the following model.

$$y_t = \theta x_t + \psi\sigma_t^2 + \varepsilon_t$$

and

$$\sigma_t^2 = \omega + \alpha(L)\varepsilon_{t-1}^2 + \beta(L)\sigma_{t-1}^2$$

Consistent estimation of θ and ψ is dependent on the correct specification of the entire model. The estimation of GARCH in Mean type of models is numerically unstable and many empirical applications have used ARCH-M type of models which are easier to estimate.

An ARCH in Mean model simply models the conditional variance as an ARCH model instead of modeling as GARCH, i.e.

$$\sigma_t^2 = \omega + \alpha(L)\varepsilon_{t-1}^2$$

Example of an ARCH(1)-M

Consider a simple version of the above model.

$$y_t = \psi\sigma_t^2 + \varepsilon_t$$

where $\varepsilon_t = v_t\sigma_t \quad v_t$ ~N(0,1)

$$\sigma_t^2 = w + \omega + \alpha\varepsilon_{t-1}^2$$

Then y_t may be expressed as

$$y_t = \psi(\omega + \alpha\varepsilon_{t-1}^2) + \varepsilon_t$$

Then the expected value of y_t is

$$E(y_t) = \psi\omega + \psi\alpha E(\varepsilon_{t-1}^2)$$

and using that $E(\varepsilon_{t-1}^2) = \omega/(1-\alpha)$then

$$E(y_t) = \psi\omega + \psi\alpha\omega/(1-\alpha)$$

Which can be viewed in finance models as the unconditional expected return for holding a risky asset.

Testing for Arch

Before attempting to estimate a GARCH model you should first check if there are ARCH effects in the residuals of the model. Clearly we should not explicitly model (and estimate) the conditional volatility of series as GARCH when there are not signs of Arch effects.

The original Lagrange Multiplier test for ARCH proposed by Engle (1982) is very simple to compute, and relatively easy to derive. Under the null hypothesis it is assumed that the model is say an AR(p) model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$$

where ε_t is a Gaussian white noise process, $\varepsilon_t|I_{t-1}$ ~N(0,σ^2)where I_t is the information set. The Alternative hypothesis is that the errors are ARCH(q).

The test for ARCH(q) effect simply consists on regressing

$$\hat{\varepsilon}_t^2 = \alpha_1\hat{\varepsilon}_{t-1}^2 + \alpha_2\hat{\varepsilon}_{t-2}^2 + ... + \alpha_q\hat{\varepsilon}_{t-q}^2 + \psi_t$$

Under the null hypothesis that $\alpha_1 = \alpha_2 = ... = \alpha_q = 0$, and TR^2 is asymptotically distributed $\chi(q)$, where T is the number of observations.

While this is the most widely used test we should be cautious in interpreting the results. If the model is misspecified it is quite likely to reject the null hypothesis simply because most of the time serial correlation in the residuals will induce serial correlation in the squared residuals.

Structural Breaks and ARCH effects

It has been shown (Diebold (1986)) that breaks in the variance, which are not taken into account by the econometrician, will look as ARCH effects when the whole sample is used. In other words, it might be that for a sub sample the unconditional variance changes from say to and then back to the previous level. In this case to model the conditional variance as an ARCH model will be the wrong thing to do. In this case, it is recommended to divide the sample and test for ARCH for the sub periods, if no ARCH effects are found for any of the sub periods but are found for the whole sample that is a clear indication of a break in the unconditional variance and not of ARCH effects. Many researchers wrongly estimated GARCH Models in many situations where there was only a change in regime. For example many papers use GARCH models to fit interest

rate series for USA when the change in the Volatility was simply a result of the different operative procedures of the Federal Reserve (a different distribution).

GARCH Effects and Sampling Frequency

It can be proved that GARCH models do not temporarily aggregate, or in other words if a model is GARCH using daily data cannot be GARCH with weakly data and so on. Given that we don't observe the data generating process in practice is very difficult to determine at which sampling frequency the data presents GARCH effects (if it has at all). Nevertheless there are some well established empirical regularities that show that the higher is the sampling frequency (say daily) the higher the GARCH effects. Weakly and every forth night data seem to also present GARCH effect. Monthly data usually does not have GARCH effects and whenever these are detected, are usually due to a structural break of the unconditional variance.

Estimating GARCH Models

Maximum likelihood Estimation with Gaussian Errors

The estimation of GARCH type models is easily done by conditional maximum likelihood.

If the model to be estimated is

$$y_t = x_t\theta + \varepsilon_t$$

Where x_t is a (row) vector of predetermined variables, which could include lagged variables, θ is a parameter vector and ε_t ~N(0,σ_t^2), where the conditional variance is assumed to be GARCH(1,1), i.e. ;

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

Then the conditional distribution of y_t is

$$f(y_t|x_t, I_{t-1}) = (2\pi\sigma_t^2)^{-.5}\exp(-.5(y_t - x_t\theta)^2/\sigma_t^2)$$

Then the conditional log likelihood is

$$\log L(\theta, \omega, \alpha, \beta|I_{t-1}) = \sum_{t=1}^{T}(-.5\log(2\pi) - .5\log(\sigma_t^2) - .5\sigma_t^{-2}(y_t - x_t\theta)^2)$$

Notice that at time 1 we need initial values for ε_0 and σ_0^2. These values are usually assumed to be the equilibrium values, that is $\sigma_0^2 = \omega/(1-\alpha-\beta)$ and $\varepsilon_0 = (\omega/(1-\alpha-\beta))^{.5}$

Maximum likelihood Estimation with non Gaussian Errors

The unconditional distribution of many financial time series seems to have fatter tails than the normal. GARCH effects may not account for this and we need to use another distribution for ε_t . A tractable distribution is the t-distribution. We proceed as before but replace the Normal density function by

$$f(\varepsilon_t) = (\Gamma[(\nu+1)/2]/\Gamma(\nu/2))((\nu-2)\pi\sigma_t^2)^{-.5}[1+(\varepsilon_t^2/(\sigma_t^2(\nu-2))]^{-(\nu+1)/2}$$

Where n is a parameter to be estimated which represents the degrees of freedom. We estimate as before numerically subject to the constraint that n is greater than 2.

Stochastic Volatility Models

A possible response to non-normality of returns conditional upon past returns is to assume that there is a random variable conditional upon which returns are normal, but this variable-which we may call stochastic volatility-is not directly observed.

A simple example of a stochastic-volatility model is the following:

$$\eta_t = \varepsilon_t e^{\alpha_t/2}, \alpha_t = \phi\alpha_{t-1} + \xi_t$$

where ε_t ~N(0,σ_ε^2), ξ_t ~N(0,σ_ξ^2), and we assume that ε_t and ξ_t are serially uncorrelated and independent of each other.

If we squared η_t, the returns equation, and take logs we can write this expression as

$$\log(\eta_t^2) = \alpha_t + \log(\varepsilon_t^2), \quad \alpha_t = \phi\alpha_{t-1} + \xi_t$$

This is in linear state-space (to be covered in the course) form except that has an error with a log χ^2 distribution instead of a normal distribution.

How to compare Between Different GARCH - Specifications

Most of the GARCH models are non-nested (they cannot be written as a restricted version of a more general process). Therefore the comparison between different GARCH Models is no straight forward.

Misspecification Tests on the standardized residuals.

We have seen above that the residuals may be written as the product of a WN and the conditional standard deviation. For example for an ARCH(1) this can be written as

$$\varepsilon_t = v_t(\omega + \alpha\varepsilon_{t-1}^2)^{1/2}$$

Therefore we can test for the existence ARCH effects in the standardized residuals

$$\hat{v}_t = \hat{\varepsilon}_t / (\hat{\omega} + \hat{\alpha}\hat{\varepsilon}_{t-1}^2)^{1/2}$$

The model that "cleans" the standardized residuals is a candidate to be the "true" model.

Some Other Ways Of Discriminating between alternative ARCH models.

We are going to present two alternative ways of discriminating between ARCH models; (i) based on the use of auxiliary regressions of the squared residuals, (ii) based in their forecasting ability.

(i) Comparison between alternative models based on the use of auxiliary regressions

Pagan and Scwhert (1989) suggest to use the following auxiliary regression as a mean of choosing between different Arch models.

$$\hat{\varepsilon}_t^2 = \alpha + \beta\hat{\sigma}_t^2 + \xi_t$$

This regress the squared residuals on the fitted variance of the alternative GARCH models. If the chosen GARCH model is appropriate to explain the conditional volatility of the series under scrutiny, you should expect α to be zero, β to be one and the fit (R^2) to be good.

Pagan and Scwhert (1989) propose to test the joint hypothesis

$H_0)\alpha = 0,\ \beta = 1$
$H_1)\alpha \neq 0,\ \beta \neq 1$

As a second step, they propose to compare the models that were not rejected on the basis of goodness of fit. The argument being, the one with better fit the better that mimics the conditional variance.

They also propose to express the previous regression in logarithms to account for scale effects and then compare the goodness of fit of this alternative auxiliary regression.

(ii) Measuring the Accuracy of Forecasts of Different Arch Models.

Hamilton (1994) propose to use the forecasting ability of the different ARCH models as a way of comparing these models. As we said before, the ARCH type of models have the property that they allow to forecast the conditional variance

of a series, therefore a criteria which may enable us to choose between different models is to choose that one that forecast better.

Various measures (*loss functions*) have been proposed for assessing the predictive accuracy of the forecasting ARCH models.

The Mean Squared Error

$$MSE = (1/T)(\sum_{t=1}^{T}(\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2)^2)$$

The Mean Absolute Error.

$$MAE = (1/T)(\sum_{t=1}^{T}|\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2|)$$

The Mean Squared Error of the log of the squared residuals.

$$[LE]^2 = (1/T)(\sum_{t=1}^{T}(\ln(\hat{\varepsilon}_t^2) - \ln(\hat{\sigma}_t^2))^2)$$

The Mean Absolute Error of the log of the squared residuals.

$$[MAE]^2 = (1/T)(\sum_{t=1}^{T}|ln(\hat{\varepsilon}_t^2) - ln(\hat{\sigma}_t^2)|)$$

For all the models we calculate the proportional improvement over a model which assumes constant variance, i.e., $\hat{\sigma}_t^2 = \hat{\sigma}^2$ (to account for scale effects). The model that provides the largest proportional improvement is the one to be preferred.

Hamilton also propose to compare the forecasting performance at different horizons (4 and 8 periods). That will slightly modify the above formulae in the following way;

The Mean Squared Error

$$MSE = (1/T)(\sum_{t=1}^{T}(\hat{\varepsilon}_{t+\tau}^2 -_\tau \hat{\sigma}_t^2)^2)$$

The Mean Absolute Error.

$$MAE = (1/T)(\sum_{t=1}^{T}|\hat{\varepsilon}_{t+\tau}^2 -_\tau \hat{\sigma}_t^2|)$$

The Mean Squared Error of the log of the squared residuals.

$$[LE]^2 = (1/T)(\sum_{t=1}^{T}(\ln(\hat{\varepsilon}_{t+\tau}^2) - \ln(\hat{\sigma}_t^2))^2)$$

The Mean Absolute Error of the log of the squared residuals.

$$[MAE]^2 = (1/T)(\sum_{t=1}^{T} |\ln(\hat{\varepsilon}_{t+\tau}^2) - \ln(\hat{\sigma}_t^2)|)$$

where ${}_\tau\hat{\sigma}_t^2$ is the forecast of the variance τ periods ahead given information at time t.

Multivariate GARCH models.

Financial market volatility moves together over time across assets and markets. Recognizing this commonality through a multivariate modeling framework leads to obvious gains in efficiency.

Following Bollerslev et al (1986) consider a multivariate extension of the GARCH(p,q) as follows:

Consider a system of n regression equations,

$$y_t = \mu + u_t$$

$$vech(H_t) = C + \sum_{i=1}^{p} A_i vech(u_{t-i}u'_{t-i}) + \sum_{i=1}^{q} B_i vech(H_{t-i})$$

where $u_t|I_{t-1}$~N(0,H_t)

In this formulation, $H_t = E(u_t u'_t|I_{t-1})$ is the $n \times n$ conditional variance matrix associated with the error vector u'_t and $vech(H_t)$ denotes the $(n(n+1))/2 \times 1$ vector of all the unique elements of H_t obtained by stacking the lower triangle of H_t[1] .

Also

- μ is $n \times 1$
- C is $(n(n+1))/2 \times 1$
- $A_1, A_2, ...A_p, B_1, ...B_q$ are $(n(n+1))/2 \times (n(n+1))/2$

Example: A Bivariate GARCH(1,1)

Since most empirical applications of the model have restricted attention to multivariate GARCH(1,1) systems. We first consider the easiest example is the simple bivariate process which depends on its conditional variance covariance matrix (H_t).

---

[1] In general if
$A = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$
Then
$vech(A) = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$
For a symmetric matrix
$B = \begin{bmatrix} a & c \\ c & d \end{bmatrix}_{n \times n}$
is standard to write the vector using either the lower or the upper triangular information
$vech(B) = \begin{bmatrix} a \\ c \\ d \end{bmatrix}_{(n \times (n+1)/2) \times 1}$

$$\begin{bmatrix} x_t \\ w_t \end{bmatrix} = \begin{bmatrix} \mu_x \\ \mu_w \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \nu_t \end{bmatrix}$$

where

$$\begin{bmatrix} V_{t-1}(x_t) \\ COV_{t-1}(x_t w_t) \\ V_{t-1}(w_t) \end{bmatrix} = \begin{bmatrix} c_x \\ c_{xw} \\ c_w \end{bmatrix} + \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & a_9 \end{bmatrix} \begin{bmatrix} (\varepsilon_{t-1})^2 \\ (\varepsilon_{t-1}\nu_{t-1}) \\ (\nu_{t-1})^2 \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & b_3 \\ b_4 & b_5 & b_6 \\ b_7 & b_8 & b_9 \end{bmatrix} \begin{bmatrix} V_{t-2}(x_{t-1}) \\ COV_{t-2}(x_{t-1}w_{t-1}) \\ V_{t-2}(w_{t-1}) \end{bmatrix}$$

Estimation of this model may be obtained by ML using numerical procedures. Two main problems with the estimation of these type of models should be apparent:

1. As its stands this model is very general, requiring a large number of parameters to be estimated. The "simplest" general model has 23 parameters.
2. We should ensure H_t to be positive definite. (see Bera and Higgins).

Various simplifications have been suggested to account for (1) and (2).

A Diagonal Vech Parameterization

(Bollerslev, Engle and Wooldridge (1982))

A natural simplification is to assume that each covariance depends only on its own past values and innovations, i.e., that A_i and B_i are diagonal. Then Each element of H_t follows an univariate GARCH model driven by the corresponding cross product $u_t u_t'$. The implied conditional covariance matrix is always positive definite if the matrices of parameters

$$\begin{bmatrix} c_x & c_{xw} \\ c_{xw} & c_w \end{bmatrix}, \begin{bmatrix} a_1 & a_5 \\ a_5 & a_9 \end{bmatrix}, \begin{bmatrix} b_1 & b_5 \\ b_5 & b_9 \end{bmatrix}$$

are all positive definite. This can be ensured by doing simple Cholesky transformations to each of these matrices.

A Quadratic Specification,

(Engle and Kroner(1995))

An alternative to the diagonal vech parameterization is to achieve a positive definite covariance matrix is the quadratic parameterization

$$H_t = C'C + A\prime u_{t-1}u'_{t-1}A + B'H_{t-1}B.$$

where C is a lower triangular with $n(n+1)/2$ parameters and A and B are square matrices with n^2 parameters each. Weak restrictions on A and B guarantee that the H_t is always positive definite.

A Constant-Correlation Specification,

(Bollerslev(1990))

This specification is similar to the diagonal specification, but imposes the restriction that the correlation between the assets is constant.

$$COV_{t-1}(x_t w_t) = \rho\sqrt{V_{t-1}(x_t)V_{t-1}(w_t)}$$

Where ρ is also estimated with the rest of the parameter set. The conditions to get a positive definite matrix are as in the diagonal case.

Stationarity and Co-persistence

Stationarity

The conditions for stationarity and moment convergence for the multivariate case are similar to those discussed in the univariate case. Specially, for the multivariate vech GARCH(1,1) model defined above, the minimum square error forecast for $vech(H_t)$ as of time $s < t$ takes the form

$$E_s(vech(H_t)) = [\sum_{k=0}^{t-s-1}(A_1 + B_1)^k]C + (A_1 + B_1)^{t-s}vech(H_s)$$

where $(A_1 + B_1)^0$ is equal to the identity matrix by definition.

Let $V\Lambda V^{-1}$ denote the Jordan decomposition of the matrix $A_1 + B_1$, so that $(A_1+B_1)^{t-s} = V\Lambda^{t-s}V^{-1}$. Thus, $E_s(vech(H_t))$ converges to the unconditional covariance matrix $(I - A_1 - B_1)^{-1}C$, for $t \to \infty$, if and only if the absolute value of the largest eigen value of $A_1 + B_1$ is strictly less than one.

Results for other multivariate formulations are scarce, a possible exception is the constant conditional correlations parameterization where the conditions for the model to be covariance stationary are simply determined by the conditions of *each* of the univariate conditional variances.

Co-Persistence in Variance

The empirical estimates for univariate and multivariate ARCH models often indicate a high degree of persistence in the forecast moments of the conditional variances, i.e. $E_s(H_t)_{ii}$, $i = 1, 2, ...N$, for $t \to \infty$. At the same time, the

commonality in volatility movements suggest that this persistence may be common across different series. More formally Bollerslev and Engle (1993) define the multivariate ARCH process to be *co-persistent* in variance if at least one of the elements in $E_s(H_t)$ is non-convergent for increasing forecasts, $t-s$, but there exists a linear combination $\gamma'\varepsilon_t$, such that for every forecast origin s, the forecasts of the corresponding future conditional variances $E_s(\gamma' H_t \gamma)$ converge to a finite limit, independent of time s information. The conditions for this to occur are presented in Bollerslev and Engle (1993) and are similar to the conditions for co-integration in the mean as developed by Engle and Granger (1987).

Multivariate GARCH-M models .

Engle and Bollerslev (1986) consider a multivariate extension of the gARCH-m as follows:

Consider a system of n regression equations,

$$y_t = BX_t + Dvech(H_t) + u_t$$

$$vech(H_t) = C + \sum_{i=1}^{p} A_i vech(u_{t-i}u'_{t-i}) + \sum_{i=1}^{q} B_i vech(H_{t-i})$$

where $u_t|I_{t-1}\tilde{}N(0, H_t)$

Where

- B is $n \times k$
- D is $n \times (n(n+1))/2$
- C is $(n(n+1))/2 \times 1$
- $A_1, A_2, ...A_p, B_1, ...B_q$ are $(n(n+1))/2 \times (n(n+1))/2$

A Bivariate GARCH-M(1,1) model

The easiest example is the simple bivariate process which depends on its conditional variance covariance matrix (H_t)

$$\begin{bmatrix} x_t \\ w_t \end{bmatrix} = \begin{bmatrix} \mu_x \\ \mu_w \end{bmatrix} + \begin{bmatrix} d_1 & d_2 & d_3 \\ d_4 & d_5 & d_6 \end{bmatrix} \begin{bmatrix} V_{t-1}(x_t) \\ COV_{t-1}(x_t w_t) \\ V_{t-1}(w_t) \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \nu_t \end{bmatrix}$$

where

$$\begin{bmatrix} V_{t-1}(x_t) \\ COV_{t-1}(x_t w_t) \\ V_{t-1}(w_t) \end{bmatrix} = \begin{bmatrix} c_x \\ c_{xw} \\ c_w \end{bmatrix} + \begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_4 & c_5 & c_6 \end{bmatrix} \begin{bmatrix} (\varepsilon_{t-1})^2 \\ (\varepsilon_{t-1}\nu_{t-1}) \\ (\nu_{t-1})^2 \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & b_3 \\ b_4 & b_5 & b_6 \\ b_4 & b_5 & b_6 \end{bmatrix} \begin{bmatrix} V_{t-2}(x_{t-1}) \\ COV_{t-2}(x_{t-1}w_{t-1}) \\ V_{t-2}(w_{t-1}) \end{bmatrix}$$

Estimation

The estimation of all the models presented above is carried out using conditional maximum likelihood estimation. The conditional log likelihood function for a single observation can be written as

$$L_t(\theta) = -(n/2)\log(2\pi) - (1/2)\log(H_t(\theta)) - (1/2)u_t(\theta)' H_t^{-1}(\theta) u_t(\theta)$$

where θ represents a vector of parameters, n represents number of equations and t represents time.

Then conditional on initial values for u_0 and H_0 , the likelihood function for the sample $1, .., T$ can be written as

$$L(\theta) = \sum_{t=1}^{T} L_t(\theta)$$

The maximization is usually achieved through numerical methods Notice that the model is highly non-linear and very unstable.

An application of the MGARCH-M model : Testing The CAPM

General issues

Three main difficulties associated with testing the CAPM

1. The CAPM is a statement about the relationships between ex ante risk premiums and betas, both of which there are not directly observable. This problem is usually dealt with by assuming that investors form rational expectations: the realized return on assets are drawings from the ex ante probability distribution of returns on those assets.

2. The Roll- critique; many assets are not measurable (human capital etc.) so tests of the CAPM have to be based on proxies for the market portfolio which only includes (a subset of) traded assets.

3. The CAPM is a single-period model. However, in order to test the model time series are frequently used. This is only a valid procedure if both risk premia and betas are stationary.

The Unconditional CAPM a short review of the literature.

Markowitz(1959) laid the ground work for the CAPM. He cast the investor's portfolio selection problem in terms of the expected return and variance of the return. He argued that investors would optimally hold a mean-variance efficient portfolio. Sharpe and Litener showed that if investors have homogeneous expectations and optimally hold mean-variance portfolios then, the market portfolio will also be mean-variance. They also assume the existence of lending and borrowing at a risk free rate of interest. For this version of the CAPM the expected return for asset j is[2]

$$E(r_j) = r_f + \beta_{jm}(E(r_m) - r_f)$$

Where r_j, r_f and r_m are the asset j, risk free and market rate of return respectively.

Defining excess returns $\tilde{r}$, we can rewrite the above expressions as

$$E(\tilde{r}_j) = \beta_{jm}(E\tilde{r}_m)$$

where $\tilde{r}_j = r_j - r_f, \quad \tilde{r}_m = r_m - r_f$

Since the risk free rate is assumed to be non-stochastic the two equations above are equivalent. Nevertheless in empirical applications r_f is typically stochastic and therefore the β' may differ.

Empirical tests of the Sharpe-Litener CAPM have focused on three implications of the last expression.

1. The intercept is zero
2. The parameter beta completely captures the cross sectional variation of expected excess returns
3. the market risk premium, $(E\ \tilde{r}_m)$ is positive.

Statistical Framework for the Sharpe-Lintner Version.

The CAPM is a single-period model and do not have time dimension. For econometric analysis of the model we need to add an assumption concerning the time series properties of returns over time. We assume that the returns are IID and jointly multivariate Normal.

Imposing the restriction that r_{jt} and r_{mt} are multivariate normal, it follows that

[2] In testing the CAPM we usually do some simplifying assumptions. One is that the risk free rate is often approximated by the treasury bill lending rate.

$$E(\tilde{r}_{jt}|\tilde{r}_{mt}) = \alpha + \beta\tilde{r}_{mt}$$

where

$\beta_j = cov(\tilde{r}_{jt}\tilde{r}_{mt})/var(\tilde{r}_{mt})$

and

$\alpha_j = E(\tilde{r}_{jt}) - \beta_j E(\tilde{r}_{mt})$

Therefore

$$\tilde{r}_{jt} = \alpha_j + \beta_j\tilde{r}_{mt} + \varepsilon_{jt} \tag{1}$$

Traditionally β has been estimated by OLS regression, meaning that the range of problems encountered with any linear regression model recur here as well. To test the CAPM is equivalent to test the restriction that $\alpha_j = 0$ and this can be done in this context by a simple F test. Alternatively we may consider the test in a multivariate context were

$$r_t = \alpha + \beta\tilde{r}_{mt} + \varepsilon_t$$

- $E(\varepsilon_t) = 0$
- $E(\varepsilon_t\varepsilon_t') = \Sigma$
- $E(r_{mt}) = \mu_m,\ \ E(r_{mt} - \mu_m)^2 = \sigma_m^2,\ \ Cov(r_{mt}, \varepsilon_t) = 0$

Where β is a vector of $N \times 1$ and ε_t are $N \times 1$ asset returns intercepts and disturbances.

We can estimate this model by maximum likelihood and test the N zero intercept restriction by a likelihood ratio test which is asymptotically distributed $\chi^2(N)$.

The Black (1972) version.

In the absence of a risk free asset Black (1972) derived a more general version of the CAPM. In this version the expected return of asset i en excess of the zero-beta return is linearly related to its beta.

Then

$$E(r_j) = r_{om} + \beta_{jm}(E(r_m) - r_{om})$$

where r is the return of the zero beta portfolio associated with m. This portfolio is defined to be the portfolio that has minimum variance of all the portfolios associated with m.

The econometric analysis of the Black version of the CAPM treats the zero-beta portfolio as an unobserved quantity. This version can be tested as a restriction on the real return market model

$$E(r_j) = \alpha_j + \beta_{jm} E(r_m)$$

and the implication of this version is $\alpha_{jm} = E(r_{om})(1 - \beta_{jm})$

This restrictions can also be tested by a likelihood ratio test.

Applied work

Most tests of the CAPM have used a time-series of monthly rates of return on common stocks listed in the New York Stock Exchange as a proxy of the market portfolio. Suppose that the objective is to explain the (excess) return of the j^{th} portfolio of assets r_{jt} . If there is only one asset per portfolio this will simply be the (excess) return on that asset. "Portfolios" can be interpreted in diverse ways. For example Harvey(1989) has r_{jt} as the (excess) return on equity in the j^{th} country. In the market model r_{jt} is related to the return on the market or aggregate portfolio, r_{mt} . The latter may be a simple average of the returns to all stocks in the economy or perhaps is a weighted average, with weights depending on the value of the portfolio accounted for each stock.

b) The second approach interprets the tests conditional on the investors' information set which is assumed to be jointly stationary and with multivariate normally distributed asset returns.

$$E(\tilde{r}_{jt}|I_{t-1}) = \beta_{jt} E(\tilde{r}_{mt}|I_{t-1}) \tag{2}$$

The Conditional CAPM

In (1) β_j was a constant equal to the ratio $cov(r_{jt} r_{mt})/var(r_{mt})$. However, in line with the distinction between conditional and unconditional moments, one might wish to consider models for r_{jt} in which the conditional density rather than the unconditional density returns is used. Let I_{t-1} be a set of conditioning variables including the past history of r_{jt} and r_{mt} . Then the conditional asset pricing model has

$$E(\tilde{r}_{jt}|I_{t-1}) = \beta_{jt} E(\tilde{r}_{mt}|I_{t-1}) \tag{2}$$

where $\beta_{jt} = cov(\tilde{r}_{jt}\tilde{r}_{mt}|I_{t-1})/var(\tilde{r}_{mt}|I_{t-1})$.

Because the coefficients of the conditional market model are functions of the conditional moments for r_{jt} and r_{mt} it is necessary to model this in some way.

A natural way to proceed is to allow the conditional mean for r_{jt} to depend on its conditional variance, as in GARCH-M models, and to subsequently model $cov(r_t|I_{t-1})$ by a multivariate GARCH.

Multivariate GARCH-M models:A CAPM with time-varying covariances

As discussed above the conditional CAPM can be written as

$$E(r_{jt}|I_{t-1}) - r_{ft-1} = \beta_{jt}[E(r_{mt}|I_{t-1}) - r_{ft-1}]$$

where

$\beta_{jt} = cov(r_{jt}r_{mt}|I_{t-1})/var(r_{mt}|I_{t-1})$.

where, because we now allow the covariance matrix of returns

$$H = \begin{bmatrix} var(r_{jt}|I_{t-1}) & cov(r_{jt}r_{mt}|I_{t-1}) \\ cov(r_{jt}r_{mt}|I_{t-1}) & var(r_{mt}|I_{t-1}) \end{bmatrix}$$

to vary over time, both the expected returns and the betas will, in general, be time varying.

This formulation of the CAPM is, however, non-operational because of the lack of an observed series for the expected market returns.

If we assume that the "market price of risk", λ is constant, where

$$\lambda = (E(r_{mt}|I_{t-1}) - r_{ft-1})/var(r_{mt}|I_{t-1}).$$

So that

$$[E(r_{jt}|I_{t-1}) - r_{ft-1}] = \lambda cov(r_{jt}, r_{mt}|I_{t-1})$$

then we can write

$$r_{jt} = r_{ft-1} + \lambda cov(r_{jt}, r_{mt}|I_{t-1}) + u_{jt}$$

Also nothing that

$$E(r_{mt}|I_{t-1}) - r_{ft-1} = \lambda var(r_{mt}|I_{t-1})$$

(since $\lambda = (E(r_{mt}|I_{t-1}) - r_{ft-1})/var(r_{mt}|I_{t-1})$)

$$r_{mt} = r_{ft-1} + \lambda var(r_{mt}|I_{t-1}) + u_{mt}$$

Where u_{jt} and u_{mt} are the innovations.

This time varying CAPM can be put into multivariate GARCH-M form as

$$y_t = b + dvech(H_t) + u_t$$

where $y_t = (r_{jt} - r_{ft-1}, r_{mt} - r_{ft-1})'$, $vech(H_t) = (var(r_{jt}|I_{t-1}), cov(r_{jt}r_{mt}|I_{t-1}), var(r_{mt}|I_{t-1}))'$, $u_t = (u_{jt}, u_{mt})'$ and

$$d = \lambda \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The zero restrictions implied by the theory may be tested (against a general model) by a likelihood ratio test which is asymptotically distributed $\chi^2(5)$. A by-product of this methodology is that plot of conditional time varying betas can easily been obtained. This plot is very informative since it may provide as with time varying beta ranking and also show how different stocks react to shocks to the economy.

SWITCHING REGIME ESTIMATION.

Introduction.

Usually in econometrics we assume that observed data has been drawn from some data generating mechanism which may be related to some specific economic theory or simply represent the true relationship between a set of variables. An underlying assumption in all econometric models is that all the observations have been drawn from the same distribution conditional on some constant parameter set. It is very unlikely that economic time series can be characterized in such a way since we expect changes in the properties of the data when there is a change in macroeconomic policy, say from free floating to target exchange rates, or even a more dramatic change from war time to peace time.

The standard econometric approach consist of trying to detect the existence of these changes in regime using different types of parameter constancy tests, and then impose dummy variables to account for these changes. By doing this, the econometrician could ensure parameter constancy within regime. Nevertheless this procedure might be very rigid and may lead to the use of models with too many dummy variables. How should we model a change in the process followed by a particular time series? Suppose that the series under scrutiny has a break in the unconditional mean at time t_1 . For data prior to t_1 we might use a model such as

$$y_t - \mu_1 = \phi(y_{t-1} - \mu_1) + \varepsilon_t \quad for\ t_1 < t$$

and for data after t_1

$$y_t - \mu_2 = \phi(y_{t-1} - \mu_2) + \varepsilon_t \quad for\ t \geq t_1$$

Even if this specification may capture the break at t_1 , is not a satisfactory time series model. For example, how are we to forecast a series described as above. Also, if the process has change in the past it could also change again. The change in regime does not need to be the outcome of a perfectly foreseeable, deterministic event. *Rather the change itself may be regarded as a random variable.* A complete time series model would therefore include a description of the probability law governing the change from μ_1 to μ_2

We might consider the process to be influenced by an unobserved random variable x_t , which is called the *state* or *regime*. This variable may take different values at date t; if $x_t = 1$, then the process is in regime 1, while $x_t = 2$ means that the process is in regime 2. Therefore we can write this model as

$$y_t - \mu_{x_t} = \phi_1(y_{t-1} - \mu_{x_{t-1}}) + \varepsilon_t$$

where the unconditional mean takes the values μ_1 when $x_t = 1$ and the value μ_2 when $x_t = 2$.

We then need a description of the time series process for the unobserved variable. Since x_t only takes discrete values we need to model this process using a discrete-valued random variable. The easiest is a Markov chain.

Properties of Markov Processes.

Let x_t be a random variable (which denotes the unobserved state of the system) that can take values 0 and 1. If the probability that x_t takes a particular value at time t, only depends on its value at $t-1$, this variable is governed by a Markov process.

$$P((x_t = i|x_{t-1} = j, x_{t-2} = k...) = P((x_t = i|x_{t-1} = j)$$

The process is summarized by the probabilities: $P(x_t = 0|x_{t-1} = 0) = q$ and $P(x_t = 1|x_{t-1} = 1) = p$.

These information is usually summarized in what is called the transition Matrix or transition probability matrix.

	0	1	(time t-1)
0	q	$(1-p)$	
1	$(1-q)$	p	
(time t)			

Autoregressive Representation of Markov Process.

$$\begin{bmatrix} 1-x_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} q & (1-p) \\ (1-q) & p \end{bmatrix} \begin{bmatrix} 1-x_t \\ x_t \end{bmatrix} + \begin{bmatrix} \zeta_{1,t+1} \\ \zeta_{2,t+1} \end{bmatrix}$$

or

$$W_{t+1} = \mathbf{P}W_t + U_{t+1}$$

where

$W_{t+1} = \begin{bmatrix} 1-x_{t+1} \\ x_{t+1} \end{bmatrix}$ and $U_{t+1} = \begin{bmatrix} \zeta_{1,t+1} \\ \zeta_{2,t+1} \end{bmatrix}$ and $E_t\zeta_{1,t+1} = 0$, $E_t\zeta_{2,t+1} = 0$, so $E_tU_{t+1} = 0$

Then it follows that W_t is a random vector that takes the value

$W_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ when $x_t = 0$

and

$W_t = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ = when $x_t = 1$

$E(W_{t+1}|x_t = i) = \begin{bmatrix} q \\ 1-q \end{bmatrix}$ when $i = 0$

$E(W_{t+1}|x_t = i) = \begin{bmatrix} 1-p \\ p \end{bmatrix}$ when $i = 1$.

The above vectors are simply column $i+1$ of the transition matrix.

Then

$$E(W_{t+1}|W_t) = \mathbf{P} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and

$$U_{t+1} = W_{t+1} - E(W_{t+1}|W_t) = W_{t+1} - \mathbf{P} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Now notice that the second row gives

$$x_{t+1} = (1-q) + (-1+p+q)x_t + \zeta_{2,t+1}$$

This expression can be recognized as an AR(1) process with constant term 1−q and autoregressive coefficient $(-1+p+q)$

The process is stationary whenever the autoregressive coefficient is smaller than 1, i.e., $(-1+p+q) < 1$,or ,$p+q < 2$.

Then the expected value of x_t is given by

$$E(x_{t+1}) = (1-q) + (-1+p+q)E(x_t)$$

or

$$E(x_t) = (1-q)/(2-p-q)$$

since $E(x_{t+1}) = E(x_t)$ for a stationary process. Also notice that the expected value may be written as

$$E(x_t) = 0.\ \ P(x_t = 0) + 1P(x_t = 1) = P(x_t = 1)$$

therefore the **unconditional probabilities** of being in state 1 and zero are

$$P(x_t = 1) = (1-q)/(2-p-q)$$

and

$$P(x_t = 0) = 1 - P(x_t = 1) = 1 - (1-q)/(2-p-q) = (1-p)/(2-p-q)$$

Conditional and Unconditional Probabilities of States 0 and 1

(An alternative derivation).

Notice that to start the Markov chain we need information of the probabilities of state 0 and 1 at time zero. This is given by the unconditional probabilities $P(x_0 = 0)$ and $P(x_0 = 1)$. To get the unconditional probabilities of the states

at time 1, you simply need to multiply the unconditional probabilities at time zero by the matrix of transition probabilities.

$$\begin{bmatrix} P(x_1 = 0) \\ P(x_1 = 1) \end{bmatrix} = \begin{bmatrix} q & (1-p) \\ (1-q) & p \end{bmatrix} \begin{bmatrix} P(x_0 = 0) \\ P(x_0 = 1) \end{bmatrix}$$

It is shown in Cox and Miller (1965) that if there exists a statistical equilibrium in which the state equilibrium is independent of the initial conditions, then the state probability π satisfies the condition that $\Pi = \mathbf{P}\Pi$, where $\mathbf{P}$ is the matrix of transition probabilities.

$$\begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix} = \begin{bmatrix} q & (1-p) \\ (1-q) & p \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix}$$

where $\pi_0 = P(x_{t-j} = 0)$ and $\pi_1 = P(x_{t-j} = 1)$ for all values of j.

Then using $\pi_0 + \pi_1 = 1$ we get the following values:

$$\pi_0 = \frac{(1-p)}{(2-p-q)} \text{ and } \pi_1 = \frac{(1-q)}{(2-p-q)}$$

where π_0 and π_1 are the equilibrium unconditional probabilities. If there exists a pair of initial values that introduce stationarity in the stochastic process, then choice of the initial value is of great importance. Note that the initial values can be the equilibrium unconditional probability.

Given the following initial values,

$$\begin{bmatrix} p^0(0) \\ p^0(1) \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix}$$

it can be shown (by multiplying n times by the transition probability matrix) that the unconditional probability vector at time n is:

$$\begin{bmatrix} p^n(0) \\ p^n(1) \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix}$$

Therefore, the distribution does not change with time and the stochastic process is always in equilibrium.

Forecasts for a Markov chain

Sometimes it is also useful to know the probability of being in state 1 (0)at time $t+n$ given that state 1 (0) prevailed at time t. A clear example is Hamilton's (1988) application of filter to test the (rational) expectational hypothesis of the term structure of interest rates which requires the forecasting of the state n periods ahead given the information of the state at time t.

. A n-periods ahead forecast for a Markov chain can be obtained simply by multiplying n times by the transition probability.

$$\begin{bmatrix} P(x_{t+n} = 0) \\ P(x_{t+n} = 1) \end{bmatrix} = \begin{bmatrix} q & (1-p) \\ (1-q) & p \end{bmatrix}^n \begin{bmatrix} P(x_t = 0) \\ P(x_t = 1) \end{bmatrix}$$

Following Cox and Miller (1965), $\mathbf{P}^n$ is derived in the following way:

1) Find the eigen-values of the transition probability Matrix.

$$\lambda_1 = 1, \quad \lambda_2 = -1 + p + q,$$

2) Find the associated eigen-vectors.

$$\begin{bmatrix} \frac{(1-p)}{(2-p-q)} \\ \frac{(1-q)}{(2-p-q)} \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

3) Express $\mathbf{P}$ as $T\Lambda T^{-1}$, where

$$T = \begin{bmatrix} \frac{(1-p)}{(2-p-q)} & -1 \\ \frac{(1-q)}{(2-p-q)} & 1 \end{bmatrix},$$

is the matrix of eigen-vectors and

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & -1+p+q \end{bmatrix},$$

a diagonal matrix of eigen-values.

4) Use the result that

$$\mathbf{P}^n = T\Lambda^n T^{-1}$$

or

$$\mathbf{P}^n = \begin{bmatrix} \frac{(1-p)}{(2-p-q)} & -1 \\ \frac{(1-q)}{(2-p-q)} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1+p+q \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -\frac{(1-q)}{(2-p-q)} & \frac{(1-p)}{(2-p-q)} \end{bmatrix}$$

	0	1	(time t)
0	$\frac{(1-p)}{(2-p-q)} + \frac{(1-q)(p+q-1)^n}{(2-p-q)}$	$\frac{(1-p)}{(2-p-q)} - \frac{(1-p)(p+q-1)^n}{(2-p-q)}$	
1	$\frac{(1-q)}{(2-p-q)} - \frac{(1-q)(p+q-1)^n}{(2-p-q)}$	$\frac{(1-q)}{(2-p-q)} + \frac{(1-p)(p+q-1)^n}{(2-p-q)}$	
(time t+n)			

Note that by making $n = 1$ in the above matrix we end with the matrix of transition probabilities, $\mathbf{P}$.

Note also that when n tends to infinity, the conditional tends to the unconditional probability. The fact that x_t has been in state 0 or 1 "infinite" number of periods ago does not provide any useful information.

In addition, we can derive the conditional expectations:

$$E(x_{t+n}|x_t = 0) = \frac{(1-q)}{(2-p-q)} - \frac{(1-q)(p+q-1)^n}{(2-p-q)}$$
$$E(x_{t+n}|x_t = 1) = \frac{(1-q)}{(2-p-q)} + \frac{(1-p)(p+q-1)^n}{(2-p-q)}$$

The expected value of x_t at time n conditional on s at time zero is:

$$E(x_{t+n}|x_t = x_t) = \frac{(1-q)}{(2-p-q)} + (x_t - \frac{(1-q)}{(2-p-q)})(p+q-1)^n$$

This result is derived assuming that x_t is observed (conditional on x_t), in most of the empirical applications we will assume that the states are unobserved and must be predicted from the realizations of y_t (the observed variable). Hamilton has done this using a non-linear filter consisting of five steps that are described below. His procedure for drawing inferences about x_t is an iterative one. Given an initial inference about x_{t-1} based on data observed through date $t-1$, iteration t produces an inference about x_t based on data observed through data t.

A Brief Description of Hamilton's Non-Linear Filter.

The filter developed by Hamilton assumes that discrete states (say high or low inflation) of the economy are not known and therefore have to be inferred from the data. He assumes that the states follow a discrete Markov process. Hamilton constructed an optimal non-linear forecast, which can be thought of as arising from a two step procedure which involves obtaining an optimal inference about the current state given the past values of the variable that is to be forecast, and then using the outcome of the filter to generate future forecasts of this variable.

Hamilton's starting point is the assumption that the economy has two possible states, let us say, state zero and state one. For the sake of simplicity, he also assumes that the unconditional mean and the variance are the only parameters that are allowed to vary between regimes. It should be noted that there is no reason to assume a priori that the other parameters will remain constant from one regime to the other, and this is something that can easily be tested using traditional procedures.

The observed variable, y_t , is assumed to follow an autoregressive process of order m, allowing the mean and the variance to vary from state 0 to state 1. This can be represented in the following way.

$$y_t - \mu_{x_t} = \phi_1(y_{t-1} - \mu_{x_{t-1}}) + \phi_2(y_{t-2} - \mu_{x_{t-2}}) + .. + \phi_m(y_{t-m} - \mu_{x_{t-m}}) + \sigma_{x_t} v_t$$

v_t is distributed N(0,1) and μ_{x_t} is parameterized as $\alpha_0 + \alpha_1 x_t$ and σ_{x_t} as $w_0 + w_1 x_t$

That is the mean is equal to α_0 in state 0 (when $x_t = 0$), and equal to $\alpha_0 + \alpha_1$ in state 1 (when $x_t = 1$) and the standard deviation is equal to w_0 in state zero and equal to $w_0 + w_1$ in state one.

The error v_t is assumed to be independent of all $x_{t-j} \geq 0$

4.2.2 Hamilton's Filter.

Step_ 1. Calculate the joint density of the m past states and the current state conditional on the information included in y_{t-1} and all past values of y, where y is the variable that is observed. This is done by using the Markov property which says that only the information of the last state is relevant.[1]

$$p(x_t, x_{t-1}, ., x_{t-m}|y_{t-1}, y_{t-2}, ., y_0) = p(x_t|x_{t-1})p(x_{t-1}, x_{t-2}, ., x_{t-m}|y_{t-1}, y_{t-2}, ., y_0)$$

$p(x_t|x_{t-1})$ is given by (4.1). As in all the subsequent steps the second term on the right-hand-side is obtained from the preceding step of the filter. In this case, $p(x_{t-1}, x_{t-2}, ..., x_{t-m}|y_{t-1}, y_{t-2}, ..., y_0)$ is known from the input to the filter, which in turn represents the result of the iteration at date $t-1$ (from step 5).

To begin with the iteration, it is necessary to assign some initial values to the parameters, and to impose some initial conditions on the Markov process. For the sake of simplicity, the unconditional distribution $p(x_{m-1}, x_{m-2}, ...x_0)$ has been chosen for the first observation.

$$p(x_{m-1}, x_{m-2}, ...x_0) = p(x_{m-1}|x_{m-2})...p(x_1|x_0)p(x_0)$$

where $p(x_0)$ are the equilibrium unconditional probabilities as defined above.

Step_ 2. Calculate the joint conditional distribution of y_t and $(x_t, x_{t-1}, ..., x_{t-m})$.

$$\begin{aligned} p(y_t, x_t, x_{t-1}, ..., x_{t-m}|y_{t-1}, y_{t-2}, ..., y_0) &= p(y_t|x_t, x_{t-1}, ..., x_{t-m}, y_{t-1}, y_{t-2}, ..., y_0). \\ & \quad p(x_t, x_{t-1}, x_{t-2}, ..., x_{t-m}|y_{t-1}, y_{t-2}, ..., y_0) \end{aligned}$$

where we assume that

$$\begin{aligned} & p(y_t|x_t, x_{t-1}, ., x_{t-m}, y_{t-1}, y_{t-2}, ., y_0) \\ = & \frac{1}{\sqrt{2\pi}(\omega_0 + \omega_1 x_t)} \exp[-\frac{1}{2[\omega_0 + \omega_1 x_t]^2}((y_t - \alpha_1 x_t - \alpha_0) \\ & -\phi_1(y_{t-1} - \alpha_1 x_{t-1} - \alpha_0) - .. - \phi_m(y_{t-m} - \alpha_1 x_{t-m} - \alpha_0))^2] \end{aligned}$$

[1] This means that the probability of the current state conditional on the previous state is equal to the probability of the current state conditional on the m past states, that is;
$P(X_t|X_{t-1}) = (X_t|X_{t-1}, ..., X_{t-m})$

Note that $p(y_t|x_t, x_{t-1}, ., x_{t-m}, y_{t-1}, y_{t-2}, ., y_0)$ involves $(x_t, x_{t-1}, ., x_{t-m})$ which is a vector which can take 2^{m+1} values.

Step_ 3. Marginalise the previous joint density with respect to the states giving the conditional density, from which the (conditional) likelihood function is calculated.

$$p(y_t|y_{t-1}, y_{t-2}, ., y_0) = \sum_{x_t=0}^{1} \sum_{x_{t-1}=0}^{1} .. \sum_{x_{t-m}=0}^{1} p(y_t, x_t, x_{t-1}, ., x_{t-m}|y_{t-1}, y_{t-2}, ., y_0)$$

Step_ 4. Combine the results from steps 2 and 3 to calculate the joint density of the state conditional on the observed current and past realizations of y

$$p(x_t, x_{t-1}, ., x_{t-m}|y_t, y_{t-1}, y_{t-2}, ., y_0) = \frac{p(y_t, x_t, x_{t-1}, ., x_{t-m}|y_{t-1}, y_{t-2}, ., y_0)}{p(y_t|y_{t-1}, y_{t-2}, ., y_0)}$$

.

Step_ 5. The desired output is then obtained from

$$p(x_t, x_{t-1}, ., x_{t-m+1}|y_t, y_{t-1}, y_{t-2}.., y_0) = \sum_{x_{t-m}=0}^{1} p(x_t, x_{t-1}, ., x_{t-m}|y_t, y_{t-1}, y_{t-2}, ., y_0)$$

The output of step 5 is used as an input to the filter in the next iteration. Note that to iterate, estimates of the parameters are required. Maximum likelihood estimates can be obtained numerically from Step 3 as a by-product of the filter, and this is the approach followed in Hamilton (1988). The sample conditional likelihood is:

$$\ln p(y_t, y_{t-1}, y_{t-2}, ., y_m)|y_{m-1}, ., y_0) = \sum_{t=m}^{T} \ln p(y_t|y_{t-1}, ., y_0).$$

which can be maximized numerically with respect to the unknown parameters $(\alpha_1, \alpha_0, p, q, \omega_0, \omega_1, \phi_1, \phi_2 ... \phi_m)$.

Notice that p and q, the parameters of the transition matrix, are also estimated by maximum likelihood. As we said in Note 1, p and q are the parameters of the transition matrix associated with remaining in the previous state.

Hamilton's filter requires the numerical optimization of a very complicated non-linear function. Cramer (1986) has pointed out that maximum likelihood estimation suffers from several specific types of failures. First, the parameter vector may change direction at ever increasing speed toward absurd values, while still increasing log likelihood at each step. Second, the iterative process may also enter a loop and keep repeating the same movements of the parameters. Third, collinearity of the data or under identification of the model can produce

a close to singular information matrix. We must bear in mind that most of these problems can occur in the examples to be analyzed as a result of the non-linear nature of Hamilton's filter.

Introduction to Bayesian Econometrics

Consider the following regression Model

$$Y = X\beta + \varepsilon, \qquad \varepsilon \tilde{} N(0, \sigma^2 I)$$

Within the Bayesian framework the parameters $\theta = \{\beta, \sigma^2\}$ are treated as random variables. This parameters have probability distributions which reflect the knowledge of the researcher, before observing the sample on Y and X, about the parameters of the model. This probability distribution, denoted as $g(\theta)$ is called a prior distribution.

Once Y is observed, the researcher revises the distribution of the parameters by combining the prior distribution with the information obtained in the sample using Bayes Theorem.

We will define the following concepts:

i) $f(Y|\theta)$ denotes the distribution of Y (from where we draw the data) given the parameters

ii) $h(\theta, Y)$ denotes the joint distribution of Y and θ.

iii) $f(Y)$ denotes the marginal distribution of Y

iv) $p(\theta|Y)$ denotes the posterior distribution of θ given Y

Then we can write the joint distribution as

$$h(\theta, Y) = f(Y|\theta)g(\theta) = p(\theta|Y)f(Y)$$

which allows to write the posterior as

$$p(\theta|Y) = \frac{f(Y|\theta)g(\theta)}{f(Y)}$$

or equivalently

$$p(\theta|Y) \propto f(Y|\theta)g(\theta)$$

Using the functional equivalence between $f(Y|\theta)$ and the likelihood $L(\theta|Y)$ we can express the posterior as

$$p(\theta|Y) \propto L(\theta|Y)g(\theta).$$

As we will see later on, the classical and the Bayesian approach are the same when the prior information is not available, that is, when the prior is diffuse or flat.

Which prior distributions should be used?

There are groups of densities that may be easier to combine with the information of the likelihood. The natural conjugate priors are priors that once they are combined with the likelihood, they produce a posterior which has the same distribution as the prior.

Example: Distribution of β assuming that σ^2 is known

Assume $\beta|\sigma^2 \tilde{} N(\beta_0, \Sigma_0)$ (a multivariate normal distribution) where β_0 and Σ_0 are known. Then the distribution can be written as

$$\begin{aligned} g(\beta|\sigma^2) &= (2\pi)^{-\frac{K}{2}} \ |\Sigma_0|^{.5} \ exp\left\{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)\right\} \\ &\propto exp\left\{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)\right\} \end{aligned}$$

where $(2\pi)^{-\frac{K}{2}} \ |\Sigma_0|^{.5}$ is a known constant.

The log likelihood

$$\begin{aligned} L(\beta|\sigma^2, Y) &= (2\pi\sigma^2)^{-\frac{T}{2}} \ exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\} \\ &\propto exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\}. \end{aligned}$$

where $(2\pi\sigma^2)^{-\frac{T}{2}}$ is a known constant.

Then the posterior is

$$\begin{aligned} p(\beta|\sigma^2, Y) &\propto g(\beta|\sigma^2)L(\beta|\sigma^2, Y) \\ &\propto exp\left\{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0) - \frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\}. \end{aligned}$$

Rearranging terms it can be shown that the posterior is also normal, therefore the normal density is the natural conjugate prior for β.

Posterior distribution of β

It can be shown that $\beta|\sigma^2, Y \tilde{} N(\beta_1, \Sigma_1)$ where

$$\begin{aligned} \beta_1 &= (\Sigma_0^{-1} + \sigma^{-2}X'X)^{-1}(\Sigma_0^{-1}\beta_0 + \sigma^{-2}X'Y) \\ &= (\Sigma_0^{-1} + \sigma^{-2}X'X)^{-1}(\Sigma_0^{-1}\beta_0 + \sigma^{-2}X'X\widehat{\beta}) \\ \Sigma_1 &= (\Sigma_0^{-1} + \sigma^{-2}X'X)^{-1} \end{aligned}$$

From the previous equation we can see that the posterior mean of β is an average of β_0 and $\widehat{\beta}$.

Example: Distribution of σ^2 assuming that β is known

The natural conjugate prior for σ^2 is the inverted Gamma distribution (the natural conjugate prior for $\frac{1}{\sigma^2}$ is the Gamma distribution)[1].

Prior of $\frac{1}{\sigma^2}|\beta \tilde{} \Gamma(\frac{\upsilon_0}{2}, \frac{\delta_0}{2})$ where υ_0 and δ_0 are known.

Then

$$\begin{aligned} g(\frac{1}{\sigma^2}|\beta) &\propto (\frac{1}{\sigma^2})^{\frac{\upsilon_0}{2}-1} \exp(-\frac{\delta_0}{2\sigma^2}) \\ & \quad and \\ L(\frac{1}{\sigma^2}|\beta, Y) &= (2\pi\sigma^2)^{-\frac{T}{2}} \; exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\} \\ &\propto (\sigma^2)^{-\frac{T}{2}} exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\}. \end{aligned}$$

The posterior, $\frac{1}{\sigma^2}|\beta, Y \tilde{} \Gamma(\frac{\upsilon_1}{2}, \frac{\delta 1}{2})$, is therefore

$$\begin{aligned} p(\frac{1}{\sigma^2}|\beta, Y) &\propto g(\frac{1}{\sigma^2}|\beta) \; L(\frac{1}{\sigma^2}|\beta, Y) \\ &\propto (\frac{1}{\sigma^2})^{\frac{\upsilon_0}{2}-1} \exp(-\frac{\delta_0}{2\sigma^2}) \; (\sigma^2)^{-\frac{T}{2}} exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\} \\ &= (\frac{1}{\sigma^2})^{\frac{\upsilon_0}{2}+\frac{T}{2}-1} \; exp\left\{-\frac{1}{2\sigma^2}(\delta_0 + (Y-X\beta)'(Y-X\beta))\right\} \\ &= (\frac{1}{\sigma^2})^{\frac{\upsilon_1}{2}-1} \; exp\left\{-\frac{\delta_1}{2\sigma^2}\right\} \\ where\ \delta_1 &= \delta_0 + (Y-X\beta)'(Y-X\beta) \text{ and } \upsilon_1 = \upsilon_0 + T. \end{aligned}$$

[1] Let $z_t \tilde{} i.i.d.N(0, \frac{1}{\delta})$, $t = 1, 2, ..., \upsilon$.

Then $W = \sum_{t=1}^{\upsilon} z_t^2 \tilde{} \Gamma(\frac{\upsilon}{2}, \frac{\delta}{2})$, where the density of the Gamma distribution is given by

$$g(W) \propto \; W^{\frac{\upsilon}{2}-1} \exp(-\frac{W\delta}{2})$$

with $E(W) = \frac{\upsilon}{\delta}$ and $V(W) = 2\frac{\upsilon}{\delta^2}$.

GIbbs- Sampling in Econometrics

Gibbs-sampling is a Markov chain Monte-Carlo method for approximating the joint and marginal distributions by sampling from conditional distributions.

Consider the following joint density

$$f(z_1, z_2, ..., z_k)$$

and that we are interested in obtaining characteristics of the marginal density

$$f(z_t) = \int ..... \int f(z_1, z_2, ..., z_k) dz_1 dz_2, ..dz_{t-1} dz_{t+1} ....., dz_k$$

such as the mean or the variance. This exercise may be, when possible, very difficult to perform

Gibbs sampling will allow me, if we are given the complete set of conditional distributions $f(z_t|z_1, z_2, ..z_{t-1}, z_{t+1}....., z_k)$, to generate a sample of $z_1, z_2, ..., z_k$ without the need of knowing the joint $f(z_1, z_2, ..., z_k)$ or the marginals $f(z_t)$.

Methodology

Given arbitrary starting values $z_2^0, ..z_t^0, z_{t+1}^0....., z_k^0$

1) Draw z_1^1 from $f(z_1|z_2^0, ..z_t^0, z_{t+1}^0....., z_k^0)$

2) Then draw z_2^1 from $f(z_2|z_1^1, z_3^0, ..z_t^0, z_{t+1}^0....., z_k^0)$

3) Then draw z_3^1 from $f(z_3|z_1^1, z_2^1, z_4^0, z_5^0....., z_k^0)$

.

.

.

k) Finally draw z_k^1 from $f(z_k|z_1^1, z_2^1, z_3^1, z_4^1, z_5^1....., z_{k-1}^1)$

Then steps 1 to k can be iterated J times to get $z_1^j, z_2^j, ..z_t^j, z_{t+1}^j....., z_k^j$, $for j = 1, 2, ..., J$.

A crucial result in the literature is that of Geman and Geman (1984) that shows that the joint and marginal distributions of $z_1^j, z_2^j, ..z_t^j, z_{t+1}^j....., z_k^j$ converge to the joint and marginal distributions of $z_1, z_2, ..z_t, z_{t+1}....., z_k$ as $J \rightarrow \infty$.

Consider $J = L + M$, then typically what is done is to use the first L simulations until the Gibbs sampler has converged and then use the remaining M simulations to approximate the empirical distribution.

Convergence of the Gibbs Sampling

The Convergence of the Gibbs sampler is a very important issue which is somehow difficult to handle. For example it is usual to plot the posterior densities over the Gibbs iterations and look for little variation in the generated distribution over the replications.

Example: A univariate Autoregression

Consider the following autoregressive model

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + e_t, \qquad e_t \tilde{} i.i.d.N(0, \sigma^2),$$

where we assume that the roots of $(1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3 - \phi_4 L^4) = 0$ lie outside the unit circle.

We can write the autoregressive model in matrix notation as

$$Y = X\beta + \varepsilon, \qquad \varepsilon \tilde{} N(0, \sigma^2 I)$$

where $\beta = [\mu,\ \phi_1,\ \phi_2,\ \phi_3,\ \phi_4]'$ and $X = [1,\ y_{t-1},\ y_{t-2},\ y_{t-3},\ y_{t-4}]$.

Conditional Distributions of β given σ^2

The *prior distribution* of β is $\beta|\sigma^2 \tilde{} N(\beta_0, \Sigma_0)_{I(s(\phi))}$, where $I(s(\phi))$ is an indicator to denote that all the roots of $(1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3 - \phi_4 L^4) = 0$ lie outside the unit circle.

The *posterior distribution* of β is $\beta|\sigma^2, Y \tilde{} N(\beta_1, \Sigma_1)_{I(s(\phi))}$,where

$$\begin{aligned} \beta_1 &= (\Sigma_0^{-1} + \sigma^{-2} X'X)^{-1}(\Sigma_0^{-1}\beta_0 + \sigma^{-2} X'Y) \\ \Sigma_1 &= (\Sigma_0^{-1} + \sigma^{-2} X'X)^{-1} \end{aligned}$$

Conditional Distributions of σ^2 given β

The *Prior* distribution of $\sigma^2|\beta \tilde{} I\Gamma(\frac{\upsilon_0}{2}, \frac{\delta_0}{2})$ where υ_0 and δ_0 are known and $I\Gamma$ denotes inverted Gamma distribution.

The *posterior distribution* of $\sigma^2|\beta \tilde{} I\Gamma(\frac{\upsilon_1}{2}, \frac{\delta_1}{2})$ *where* $\delta_1 = \delta_0 + (Y - X\beta)'(Y - X\beta)$ and $\upsilon_1 = \upsilon_0 + T$.

The Gibbs Sampling procedure consists of the following steps.

- Start the iteration of the Gibbs Sampling

To start the iteration we use an arbitrary starting value $\sigma^2 = \left\{\sigma^2\right\}^0$

- Iterate the following steps $j = L + M$ times

1) Conditional on $\sigma^2 = \left\{\sigma^2\right\}^{j-1}$, that is, the value generated in the previous iteration, we generate β^j from the posterior distribution of β.

2) Conditional on $\beta = \beta^j$,that is, the value β generated in 1), we generate $\left\{\sigma^2\right\}^j$ from the posterior distribution of σ^2.

3) Set $j = j + 1$.

In generating β we employ rejection sampling to ensure that all the roots are outside the unit circle (we discard the draws that do not satisfy this condition).

As a result of this procedure we generate the following sets of values

$$\beta^1, \beta^2, ..., \beta^{L+M},$$
$$\left\{\sigma^2\right\}^1, \left\{\sigma^2\right\}^2, ..., \left\{\sigma^2\right\}^{L+M}.$$

We discard the first L generated values to ensure convergence of the Gibbs-Sampler and then use the following M values to make inferences about β and σ^2. The remaining M values provide us with the Joint and the Marginal distribution.

Markov- Switching and Gibbs Sampling

We have shown that when estimationg M-S models we treat parameters of the model depending on an unobserved sate variable. We typically estimate the models and make inferences on the unobserved variables conditional on the parameters (estimates) of the model. The Bayesian approach treats both the parameters of the model and the Markov switching variables as random variables. Then the inference about the states of the economy (denoted as $\widetilde{S}_T = S_1, S_2, ...S_T$) is based on the joint distribution of the sates and the parameters of the model.

EXAMPLE:

Consider the following model

$$\begin{aligned} y_t &= \mu_{S_t} + \varepsilon_t \qquad \varepsilon_t \tilde{} N(0, \sigma^2_{S_t}) \\ \mu_{S_t} &= \mu_0 + \mu_1 S_t, \qquad \mu_1 > 0 \\ \sigma^2_{S_t} &= \sigma^2_0(1 - S_t) + \sigma^2_1 S_t \\ &= \sigma^2_0(1 + h_1 S_t), \qquad h_1 > -1 \end{aligned}$$

$$\begin{aligned} P(S_t &= 0|S_{t-1} = 0) = q \\ P(S_t &= 1|S_{t-1} = 1) = p \end{aligned}$$

The Bayesian approach will entail the inference about $T+6$ random variables: $\{S_1, S_2, ..., S_T, \mu_0, \mu_1, \sigma^2_0, \sigma^2_1, p, q\}$.

We need to derive the joint posterior distribution

$$\begin{aligned} g(\widetilde{S}_T, \mu_0, \mu_1, \sigma^2_0, \sigma^2_1, p, q|\widetilde{y}_T) &= g(\mu_0, \mu_1, \sigma^2_0, \sigma^2_1, p, q|\widetilde{y}_T, \widetilde{S}_T).g(\widetilde{S}_T|\widetilde{y}_T) \\ &= g(\mu_0, \mu_1, \sigma^2_0, \sigma^2_1|\widetilde{y}_T, \widetilde{S}_T).g(p, q|\widetilde{y}_T, \widetilde{S}_T).g(\widetilde{S}_T|\widetilde{y}_T) \\ &= g(\mu_0, \mu_1, \sigma^2_0, \sigma^2_1|\widetilde{y}_T, \widetilde{S}_T).g(p, q|\widetilde{S}_T).g(\widetilde{S}_T|\widetilde{y}_T) \end{aligned}$$

We assume that conditional on $\widetilde{S}_T$, p and q are independent of both other parameters of the model and of the data. Notice that conditional on $\widetilde{S}_T$, the expression $y_t = \mu_{S_t} + \varepsilon_t$, $\varepsilon_t \tilde{} N(0, \sigma^2_{S_t})$ is simply a regression model with a known dummy.

The Gibbs Sampling estimation procedure

Using arbitrary starting values we repeat the following steps.

1) Generate S_t from $g(S_t|\widetilde{S}_{\neq t}, \mu_0, \mu_1, \sigma^2_0, \sigma^2_1, p, q, \widetilde{y}_T)$, or generate the whole block $\widetilde{S}_T$ from $g(\widetilde{S}_T|\mu_0, \mu_1, \sigma^2_0, \sigma^2_1, p, q, \widetilde{y}_T)$

2) Generate the transition probabilities p and q from $g(p, q|\widetilde{S}_T)$.

3) Generate $\mu_0, \mu_1, \sigma^2_0, \sigma^2_1$, from $g(\mu_0, \mu_1, \sigma^2_0, \sigma^2_1|\widetilde{y}_T, \widetilde{S}_T)$.

Step 1):Single move Gibbs Sampling - Generate S_t from $g(S_t|\widetilde{S}_{\neq t}, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2, p, q, \widetilde{y}_T)$

Suppressing the conditioning in the parameters we can write

$$\begin{aligned}
g(S_t|\widetilde{S}_{\neq t}, \widetilde{y}_T) &= g(S_t|\widetilde{S}_{\neq t}, \widetilde{y}_t, y_{t+1}, y_{t+2}, ..., y_T) \\
&= \frac{g(S_t, y_{t+1}, y_{t+2}, ..., y_T|\widetilde{S}_{\neq t}, \widetilde{y}_t)}{g(y_{t+1}, y_{t+2}, ..., y_T|\widetilde{S}_{\neq t}, \widetilde{y}_t)} \\
&= \frac{g(S_t|\widetilde{S}_{\neq t}, \widetilde{y}_t)g(y_{t+1}, y_{t+2}, ..., y_T|\widetilde{S}_{\neq t}, \widetilde{y}_t)}{g(y_{t+1}, y_{t+2}, ..., y_T|\widetilde{S}_{\neq t}, \widetilde{y}_t)} \\
&= g(S_t|\widetilde{S}_{\neq t}, \widetilde{y}_t) \\
&= g(S_t|\widetilde{S}_{t-1}, S_{t+1}, ..., S_T, \widetilde{y}_{t-1}, y_t) \\
&= \frac{g(S_t, S_{t+1}, ..., S_T, y_t|\widetilde{S}_{t-1}, \widetilde{y}_{t-1})}{g(S_{t+1}, ..., S_T, y_t|\widetilde{S}_{t-1}, \widetilde{y}_{t-1})} \\
&\propto g(S_t, S_{t+1}, ..., S_T, y_t|\widetilde{S}_{t-1}, \widetilde{y}_{t-1}) \\
&= g(S_t|\widetilde{S}_{t-1}, \widetilde{y}_{t-1})g(S_{t+1}, ..., S_T, y_t|S_t, \widetilde{S}_{t-1}, \widetilde{y}_{t-1}) \\
&= g(S_t|S_{t-1})g(S_{t+1}, ..., S_T, y_t|S_t, \widetilde{S}_{t-1}, \widetilde{y}_{t-1})
\end{aligned}$$

Notice that

$$\begin{aligned}
g(S_{t+1}, ..., S_T, y_t|S_t, \widetilde{S}_{t-1}, \widetilde{y}_{t-1}) &= g(y_t|S_t, \widetilde{S}_{t-1}, S_{t+1}, ..., S_T, \widetilde{y}_{t-1})g(S_{t+1}, ..., S_{Tt}|S_t, \widetilde{S}_{t-1}, \widetilde{y}_{t-1}, y_t) \\
&= g(y_t|S_t)g(S_{t+1}|S_t, \widetilde{S}_{t-1}, \widetilde{y}_{t-1})g(S_{t+2}, ..., S_T, |S_{t+1}, S_t, \widetilde{S}_{t-1}, \widetilde{y}_{t-1}, y_t) \\
&= g(y_t|S_t)g(S_{t+1}|S_t)g(S_{t+2}, ..., S_T|S_{t+1}) \\
&\propto g(y_t|S_t)g(S_{t+1}|S_t)
\end{aligned}$$

Then, we can write

$$g(S_t|\widetilde{S}_{\neq t}, \widetilde{y}_T) \propto g(S_t|S_{t-1})g(y_t|S_t)g(S_{t+1}|S_t)$$

where $g(S_t|S_{t-1})$ and $g(S_{t+1}|S_t)$ are given by the transition probabilities and

$$g(y_t|S_t) = \frac{1}{\sqrt{2\pi\sigma_{s_t}^2}} \exp\{-\frac{1}{2\sigma_{s_t}^2}(y_t - \mu_{s_t})^2\}$$

We can then calculate

$$P(S_t = j|\widetilde{S}_{\neq t}, \widetilde{y}_T) = \frac{g(S_t = j|\widetilde{S}_{\neq t}, \widetilde{y}_T)}{\sum_{j=0}^{1} g(S_t = j|\widetilde{S}_{\neq t}, \widetilde{y}_T)}$$

We generate S_t using a uniform distribution between 0 and 1. If the generated number is less or equal than $P(S_t = j|\widetilde{S}_{\neq t}, \widetilde{y}_T)$, we set the value of S_t to zero, otherwise we set the value equal to one.

Step 1):Multimove Gibbs Sampling - Generate $\widetilde{S}_t$ from $g(\widetilde{S}_T|\mu_0, \mu_1, \sigma_0^2, \sigma_1^2, p, q, \widetilde{y}_T)$

Suppressing the conditioning in the parameters we can write

$$\begin{aligned}
g(\widetilde{S}_T|\widetilde{y}_T) &= g(S_1, S_2, ..., S_T, |\widetilde{y}_T) \\
&= g(S_T, |\widetilde{y}_T) g(S_1, S_2, ..., S_{T-1}, |S_T, \widetilde{y}_T) \\
&= g(S_T, |\widetilde{y}_T) g(S_{T-1}, |S_T, \widetilde{y}_T) g(S_1, S_2, ., S_{T-2}, |S_{T-1}, S_T, \widetilde{y}_T) \\
&= g(S_T, |\widetilde{y}_T) g(S_{T-1}, |S_T, \widetilde{y}_T) g(S_{T-2}, |S_{T-1}, S_T, \widetilde{y}_T) .. \\
&\quad .....g(S_1, |S_2, ., S_{T-2}, S_{T-1}, S_T, \widetilde{y}_T) ... \\
&= g(S_T, |\widetilde{y}_T) g(S_{T-1}, |S_T, \widetilde{y}_{T-1}) g(S_{T-2}, |S_{T-1}, \widetilde{y}_{T-2}) .. g(S_1, |S_2, y_1) \\
&= g(S_T, |\widetilde{y}_T) \prod_{t=1}^{T-1} g(S_t, |S_{t+1}, \widetilde{y}_t).
\end{aligned}$$

The derivation is based on the Markov property that states that to makes inference about S_t conditional on S_{t+1} the variables S_{t+2},....S_T, y_{t+1},....y_T have no information beyond that contained in S_{t+1}.

Then we proceed in the following way: we first generate $\widetilde{S}_T$ conditional on $\widetilde{y}_T$ and then, for the other values of $t = T-1, t-2, ..., 1$, we generate S_t conditional on y_t and the generated $t+1$.

We can carry out this using the following steps:

Step 1 Run the Hamilton filter to get $g(S_t|\widetilde{y}_t)$. The last iteration of the filter provides $g(S_T|\widetilde{y}_T)$ that is used to generate S_T.

Step 2 Generate S_t conditional on S_{t+1} and $\widetilde{y}_t$, for $t = T-1, t-2, ..., 1$, form $g(S_t|S_{t+1}, \widetilde{y}_t)$ using the fact that

$$\begin{aligned}
g(S_t|S_{t+1}, \widetilde{y}_t) &= \frac{g(S_t, S_{t+1}|\widetilde{y}_t)}{g(S_{t+1}|\widetilde{y}_t)} \\
&= \frac{g(S_{t+1}|S_t, \widetilde{y}_t).g(S_t|\widetilde{y}_t)}{g(S_{t+1}|\widetilde{y}_t)} \\
&= \frac{g(S_{t+1}|S_t).g(S_t|\widetilde{y}_t)}{g(S_{t+1}|\widetilde{y}_t)} \\
&\propto g(S_{t+1}|S_t).g(S_t|\widetilde{y}_t)
\end{aligned}$$

Then we calculate

$$P(S_t = 1|S_{t+1}, \widetilde{y}_T) = \frac{g(S_{t+1}|S_t = 1)g(S_t|\widetilde{y}_t)}{\sum_{j=0}^{1} g(S_{t+1}|S_t = j)g(S_t = j|\widetilde{y}_t)}$$

We generate S_t using a uniform distribution between 0 and 1. If the generated number is less or equal than $P(S_t = 1|S_{t+1}, \widetilde{y}_T)$, we set the value of S_t to zero, otherwise we set the value equal to one.

Generating Transition Probabilities p and q, conditional on $\widetilde{S}_T$

Conditional on $\widetilde{S}_T$, p and q are independent of the data set $\widetilde{y}_T$, and the other parameters of the models[2].

Prior Distribution

$p \tilde{} beta(u_{11}, u_{10})$,
$q \tilde{} beta(u_{00}, u_{01})$,

with $g(p, q) \propto p^{u_{11}-1}(1-p)^{u_{10}-1}q^{u_{00}-1}(1-q)^{u_{01}-1}$, where the $u's$ are known hyper parameters of the priors.

The likelihood function

The likelihood function is given by

$L(p, q|\widetilde{S}_T) = p^{n_{11}}(1-p)^{n_{10}}q^{n_{00}}(1-q)^{n_{01}}$ where n_{ij} refers to the number of transitions from i to j which can be counted from $\widetilde{S}_T$.

Posterior distribution

$$\begin{aligned} g(p, q|\widetilde{S}_T) &= g(p, q)L(p, q|\widetilde{S}_T) \\ &\propto p^{u_{11}-1}(1-p)^{u_{10}-1}q^{u_{00}-1}(1-q)^{u_{01}-1}p^{n_{11}}(1-p)^{n_{10}}q^{n_{00}}(1-q)^{n_{01}} \\ &= p^{u_{11}+n_{11}-1}(1-p)^{u_{10}+n_{10}-1}q^{u_{00}+n_{00}-1}(1-q)^{u_{01}+n_{01}-1}, \end{aligned}$$

[2] A beta distribution denoted by $z \tilde{} beta(z|\alpha_0, \alpha_1)$, is dependent on two hyperparameters $\alpha_0, \alpha_1 > 0$ and has density given by

$$\begin{aligned} g(z|\alpha_0, \alpha_1) &\propto z^{\alpha_0-1}(1-z)^{\alpha_1-1} \, for \; 0 < z < 1 \\ &= 0 \qquad\qquad\qquad\quad otherwise, \end{aligned}$$

with $E(z) = \frac{\alpha_0}{\alpha_0+\alpha_1}$ and $Var(z) = \frac{\alpha_0\alpha_1}{(\alpha_0+\alpha_1)^2(\alpha_0+\alpha_1+1)}$

then,
$p|\widetilde{S}_T \tilde{} beta(u_{11+}n_{11}, u_{10}+n_{10})$,
$q|\widetilde{S}_T \tilde{} beta(u_{00}+n_{00}, u_{01}+n_{01})$.

Generating μ_0, μ_1,conditional on $\sigma_0^2, \sigma_1^2, \widetilde{y}_T$ and $\widetilde{S}_T$

Given

$$\begin{aligned} y_t &= \mu_0 + \mu_1 S_t + \varepsilon_t \qquad \varepsilon_t \tilde{} N(0, \sigma^2_{S_t}) \\ &\quad \text{we can write} \\ y_t^* &= \mu_0 x_{0t} + \mu_1 x_{0t} + \upsilon_t \qquad \upsilon_t \tilde{} N(0,1) \\ \text{where } y_t^* &= \frac{y_t}{\sigma_{S_t}},\ x_{0t} = \frac{1}{\sigma_{S_t}} \text{ and } x_{1t} = \frac{S_t}{\sigma_{S_t}} \end{aligned}$$

Prior Distribution

We can write the model in matrix notation as

$$Y = X\mu + V, \qquad V \tilde{} N(0, I)$$

then if we assume a normal prior $\mu|\sigma_0^2, \sigma_1^2 \tilde{} N(b_0, B_0)$, where b_0, B_0 are given.

Posterior distribution

$\mu|\sigma_0^2, \sigma_1^2, \widetilde{S}_T, \widetilde{y}_T \tilde{} N(b_1, B_1)$, where b_1, B_1 are

$$\begin{aligned} b_1 &= (B_0^{-1} + X'X)^{-1}(B_0^{-1}b_0 + X'Y) \\ B_1 &= (B_0^{-1} + X'X)^{-1} \end{aligned}$$

to constrain $\mu_1 > 0$, we discard the draws where this condition is not satisfied.

Generating σ_0^2, σ_1^2,conditional on $\mu_0, \mu_1, \widetilde{y}_T$ and $\widetilde{S}_T$

Given

$$\begin{aligned} \sigma^2_{S_t} &= \sigma_0^2(1 + h_1 S_t), \qquad h_1 > -1 \\ &\quad where \\ \sigma_1^2 &= \sigma_0^2(1 + h_1). \end{aligned}$$

we can first generate σ_0^2 conditional on h_1, and then generate $(1 + h_1)$ conditional on σ_0^2.

Generating σ_0^2 conditional on h_1

We divide both sides of y_t by $\sqrt{(1+h_1 S_t)}$:

$$\begin{aligned} y_t^{**} &= \mu_0 x_{0t}^* + \mu_1 x_{0t}^* + \upsilon_t^* \qquad \upsilon_t^* \tilde{} N(0,\sigma_0^2) \\ \text{where } y_t^{**} &= \frac{y_t}{\sqrt{(1+h_1 S_t)}},\ x_{0t}^* = \frac{1}{\sqrt{(1+h_1 S_t)}}, \\ x_{1t}^* &= \frac{S_t}{\sqrt{(1+h_1 S_t)}} \text{ and } \upsilon_t^* = \frac{\varepsilon_t}{\sqrt{(1+h_1 S_t)}} \end{aligned}$$

then:

The Prior distribution of $\sigma_0^2|\mu_0,\mu_1,h_1 \tilde{} I\Gamma(\frac{\upsilon_0}{2},\frac{\delta_0}{2})$ where υ_0 and δ_0 are known and $I\Gamma$ denotes inverted Gamma distribution.

The posterior distribution of $\sigma_0^2|\mu_0,\mu_1,h_1,\widetilde{S}_T,\widetilde{y}_T \tilde{} I\Gamma(\frac{\upsilon_1}{2},\frac{\delta_1}{2})$ *where* $\delta_1 = \delta_0 + \sum_{t=1}^{T}(y_t^{**} - \mu_0 x_{0t}^* - \mu_1 x_{0t}^*)^2$ and $\upsilon_1 = \upsilon_0 + T$.

Generating h_1 conditional on σ_0^2

We divide both sides of y_t by σ_0 :

$$\begin{aligned} y_t^{***} &= \mu_0 x_{0t}^{**} + \mu_1 x_{0t}^{**} + \upsilon_t^{**} \qquad \upsilon_t^{**} \tilde{} N(0,(1+h_1 S_t)) \\ \text{where } y_t^{**} &= \frac{y_t}{\sigma_0},\ x_{0t}^{**} = \frac{1}{\sigma_0},\ x_{1t}^{**} = \frac{S_t}{\sigma_0} \text{ and } \upsilon_t^{**} = \frac{\varepsilon_t}{\sigma_0} \end{aligned}$$

then:

The Prior distribution of $h_1|\mu_0,\mu_1,\sigma_0^2 \tilde{} I\Gamma(\frac{\upsilon_2}{2},\frac{\delta_2}{2})$ where υ_2 and δ_2 are known and $I\Gamma$ denotes inverted Gamma distribution.

The posterior distribution of $h_1|\mu_0,\mu_1,\sigma_0^2,\widetilde{S}_T,\widetilde{y}_T \tilde{} I\Gamma(\frac{\upsilon_3}{2},\frac{\delta_3}{2})$ *where* $\delta_3 = \delta_2 + \sum_{t=1}^{N_1}(y_t^{***} - \mu_0 x_{0t}^* - \mu_1 x_{0t}^*)^2$ and $\upsilon_3 = \upsilon_2 + T$, where N_1 is the number of times $S_t = 1$..