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

$$\text{i)}\phi_1 = .5$$

$$\text{ii)}\phi_1 = -.5$$

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.

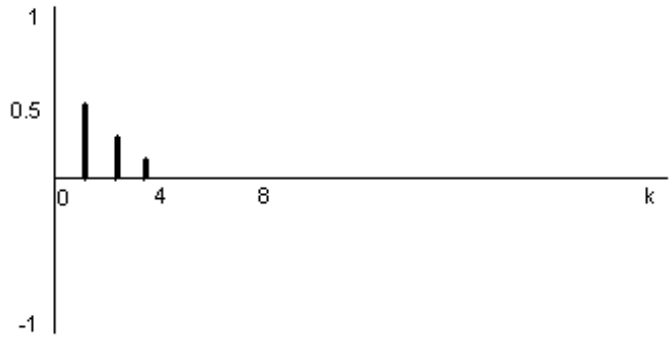

Figure 1:

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

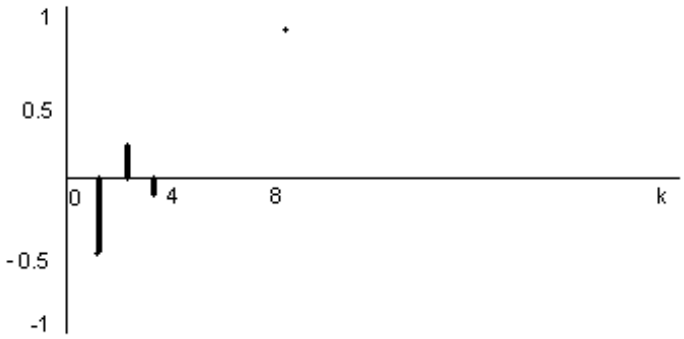

Figure 2:

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

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

$\rho_1 = \phi_1 = \phi_{11}$ for $k = 1$.
$\phi_{ii} = 0$ for $k > 1$.

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

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

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

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

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 .

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

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.

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

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

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

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

The ARMA(1,1) Process

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

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

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

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

(ii) The mean square error associated with a linear projection is given by $E(Y_{t+1} - \alpha\prime X_t)^2 = E(Y_{t+1})^2 - E(Y_{t+1}X_t)(E(X_tX_t'))^{-1}E(X_tY_{t+1})$ (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

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

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

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 $E(W_2W_1) = 0$ (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 $E(W_2W_2') = D_{22} = \Omega_{22}$-$\Omega_{21}\Omega_{11}^{-1}\Omega_{12}$

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.

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