Notes on Time Series Analysis*

Luca Sala†

Department of Economics and IGIER, Università Bocconi

May 2013

*These notes have been prepared for the course Advanced Econometrics, DES, Università Bocconi. The first draft was prepared in the Spring 2009. Please, let me know of any typos or imprecisions in them.

†Dept. of Economics and IGIER, Università Bocconi, Via Roentgen 1, 20136 Milan, Italy. E-mail: luca.sala@unibocconi.it

Stochastic processes:

A stochastic process is a double-infinite sequence of random variables:

$$\{..., X_{-k}, ..., X_0, ...X_k\}$$

or more compactly, $\{X_t\}_{t=-\infty}^{\infty}$.

Given a stochastic process, our aim is to characterize the joint distribution of the random variables in $\{X_t\}_{t=-\infty}^{\infty}$.

We can for example, define the first two moments of the random variables in $\{X_t\}_{t=-\infty}^{\infty}$.

The expectation $E(X)$ is a vector:

$$[..., E(X_{-k}), ..., E(X_0), ...E(X_k)]'$$

in which each element $E(X_k)$ is:

$$E(X_k) = \int X_k(w) f_{X_k}(w) dw$$

The autocovariance function is: $Cov(X_k, X_h) = E(X_k X_h) - E(X_k)E(X_h)$ for any k, h.

(Often, we will deal with zero-mean stochastic processes, so that $Cov(X_k, X_h) = E(X_k X_h)$).

At this level of generality, the expectation is a double infinite vector and the autocovariance function is composed by an even larger number of elements.

A realization of a stochastic process is a double infinite sequence of numbers: $\{X_t(\omega)\}_{t=-\infty}^{\infty}$, in which ω represents the event that has realized.

If we want to estimate the elements of the vector $E(X)$ we could (under some additional assumptions) replace the expectation $E(X_k)$ with the sample mean: $\frac{1}{N}\sum_{i=0}^{N} X_k(\omega_i)$ (note that each realization ω is a sequence of numbers) and hope that in some sense, $\frac{1}{N}\sum_{i=0}^{N} X_k(\omega_i) \to E(X_k)$ as $N \to \infty$ (this is the "law of large numbers").

The problem is that we typically observe only one realization of the stochastic process $\{X\}_{t=-\infty}^{\infty}$ (we only observe one path for say, the unemployment rate in the last 50 years...)

In practice, this means that inference in a time-series context is performed with only 1 observation of the underlying stochastic mechanism.

Is it meaningful? We need additional assumptions.

Stationarity

A stochastic process is *covariance (weakly) stationary*, if

$$\begin{aligned} E(X_k) &= \mu \\ Cov(X_t, X_{t+h}) &= \gamma_h \end{aligned}$$

Note that this assumes that γ_h exists and it is finite.

A stochastic process is *(strongly) stationary* if the joint distribution of any set of elements in $\{X\}$ depends only on their relative position and not on t. That is, $f(X_4, X_8) = f(X_6, X_{10})$ and so on...

With stationarity, we reduce the dimensionality of $E(X)$ and of $Cov(X_k, X_h)$ significantly! There are much less elements characterizing the first and second moments of our stochastic process.

Ergodicity.

We also need an additional condition, that guarantees that as we collect more and more observations "we keep on learning something new" on the moments of $\{X\}$. We need a condition that says that the process keeps on assuming different values from the support of its distribution.

I am not very formal here, but I hope intuition is clear...

Think of a process like this: $X \sim N(\mu, \sigma^2)$ for which we want to estimate the mean μ. At $t = -\infty$ a value is drawn and since then X remains constant at the value drawn. This process is not ergodic: as we collect more and more observations, we do not learn anything new on μ. The process remains stuck at the particular value of X that realized.

Ergodicity guarantees that:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} x_t \to \mu$$

(note that I am not very specific on what the symbol $\to$ means).

If we have stationarity, we can define the autocovariance function as depending only on s:

$$\gamma_s = Cov(X_t, X_{t+s})$$

Notice that $\gamma_s = \gamma_{-s}$.

The autocorrelation function is: $\rho_s = \gamma_s/\gamma_0$.

If in addition, we have ergodicity, we can define estimators for μ, γ_s and ρ_s by replacing population moments with sample counterparts:

$$\begin{aligned}
\hat{\mu} &= \frac{1}{T}\sum_{t=1}^{T} x_t \\
\hat{\gamma}_s &= \frac{1}{T}\sum_{t=s+1}^{T} (x_t - \hat{\mu})(x_{t-s} - \hat{\mu}) \\
\hat{\rho}_s &= \hat{\gamma}_s/\hat{\gamma}_0
\end{aligned}$$

White noise

A white noise is a stochastic process such that:

$$\begin{aligned} E(\epsilon_t) &= 0 \\ E(\epsilon_t^2) &= \sigma_\epsilon^2 \\ Cov(\epsilon_t, \epsilon_{t+s}) &= 0, \forall s \neq t \end{aligned}$$

We only have assumptions on the unconditional properties of the process, but the conditional first and second moments are left unspecified.

Sometimes, we may need to strengthen the above conditions by defining a white noise sequence of *independent and identically distributed* (i.i.d.) variables. A standard distributional assumption is: $\epsilon_t \sim iid\ N(0, \sigma_\epsilon^2)$. Under this strong definition, we are saying not only that the various ϵ are uncorrelated, but they are also independent (recall that under normality, lack of correlation $\Longleftrightarrow$ independence).

ARMA models

By taking linear combinations of white noises we can build a large class of interesting stochastic processes: the ARMA models.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$$

Special cases are:

$$\begin{aligned} AR(1) &: \quad x_t = \phi_1 x_{t-1} + \epsilon_t \\ MA(1) &: \quad x_t = \epsilon_t + \theta_1 \epsilon_{t-1} \\ ARMA(1,1) &: \quad x_t = \phi_1 x_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1} \end{aligned}$$

Lag operators:

We can simplify notation by defining the *lag operator*. The lag operator, L, is such that:

$Lx_t = x_{t-1}$
$L^2 x_t = LLx_t = Lx_{t-1} = x_{t-2}$
$L^j x_t = x_{t-j}$
$L^{-j} x_t = x_{t+j}$
$L(x_t + y_t) = x_{t-1} + y_{t-1}$
$L\beta x_t = \beta L x_t = \beta x_{t-1}$

In summary, it has the same properties of multiplication.
$ARMA$ can be written in simple form thanks to the lag operator:

The $ARMA(p,q)$

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$$

can be written as:

$$(1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) x_t = (1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q) \epsilon_t$$

or more compactly

$$\phi(L) x_t = \theta(L) \epsilon_t$$

Lag operators are very useful as we can apply to them the standard rules of algebra:

a) we can multiply them:

$$a(L)b(L) = (a_0 + a_1 L + ...)(b_0 + b_1 L + ...) = a_0 b_0 + (a_0 b_1 + b_0 a_1) L + ...$$

b) $a(L)b(L) = b(L)a(L)$
c) $a(L)^2 = a(L)a(L)$
d) we can factorize them:

$$\begin{aligned} a(L) &= (1 - \lambda_1 L)(1 - \lambda_2 L)... \text{ and (provided } |\lambda_1| < 1 \text{ and } |\lambda_2| < 1): \\ a(L)^{-1} &= (1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1} = \sum_{j=0}^{\infty} \lambda_1^j L^j \sum_{j=0}^{\infty} \lambda_2^j L^j \end{aligned}$$

An important thing to notice is that:

ARMA models are not unique.

The same process x_t can be represented by different ARMA processes.

As an example, think of an $AR(1)$:

$$x_t = \phi_1 x_{t-1} + \epsilon_t$$

substitute once backwards: $x_t = \phi_1(\phi_1 x_{t-2} + \epsilon_{t-1}) + \epsilon_t = \phi_1^2 x_{t-2} + \phi_1 \epsilon_{t-1} + \epsilon_t$
An $AR(1)$ can be re-written as an $ARMA(2,1)$, or if one keeps on substituting, as an $ARMA(k, k-1)$.

It is very important to stress that:

the autocovariance function is unique

An autocovariance function $\{\gamma_s\}$ identifies uniquely a stochastic process and a stochastic process has one and only one $\{\gamma_s\}$.

This means that the $AR(1)$ and the $ARMA(2,1)$ have to have the same $\{\gamma_s\}$.

Let us see how to solve the ARMa difference equations.

From an AR(1) to a MA(∞)

An AR(1):

$$x_t = \phi_1 x_{t-1} + \epsilon_t$$

is a stochastic difference equation, in which ϵ_t is the forcing term.

A solution to this equation is obtained by substituting recursively backwards k times:

$$x_t = \phi_1^k x_{t-k} + \sum_{j=0}^{k-1} \phi_1^j \epsilon_{t-j}$$

If $|\phi_1| < 1$, then $\lim_{k\to\infty} \phi_1^k x_{t-k} = 0$ and

$$x_t = \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}$$

Can we use the lag operator to do the same operation? The AR(1) can be written as

$$(1 - \phi_1 L) x_t = \epsilon_t$$

We can "invert" $(1 - \phi_1 L)$ to obtain

$$x_t = (1 - \phi_1 L)^{-1} \epsilon_t$$

We can think of $(1-z)^{-1}$ as the limit of a geometric sum $(1+z+z^2+z^3+...)$, with $|z| < 1$

In the same way, we obtain

$$(1 - \phi_1 L)^{-1} = (1 + \phi_1 L + \phi_1^2 L^2 + \phi_1^3 L^3 + ...)$$

$x_t = (1 - \phi_1 L)^{-1} \epsilon_t$ is then equal to

$$x_t = \sum_{j=0}^{\infty} \phi^j L^j \epsilon_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$$

as above.

We could also understand the properties of $(1-\phi L)^{-1}$ using the definition of the inverse: $(1-\phi L)^{-1}(1-\phi L)=1$. If $(1-\phi L)^{-1}=(1+\phi L+\phi^2 L^2+...)$, then the property of the inverse given before is satisfied.

If $|\phi|>1$ we cannot "invert" $(1-\phi_1 L)$ as the summation in terms of present and past ϵ_t won't have a well defined limit! Same for the backwards substitution: $\lim_{k\to\infty}\phi_1^k x_{t-k}$ is not zero.

This does not mean that a stationary process cannot be represented as an AR(1) with $|\phi|>1$ (these processes are called non-causal AR, as the solution for x_t will be a function of future ϵ_t and past ϵ_t will not "cause" x_t. This is not very intuitive... Moreover, a non-causal AR can always be rewritten as a causal one (recall that a stochastic process has one and only one autocovariance function, but that $ARMA$ are not unique)).

From AR(p) to MA(∞)

Backwards substitution is too complicated to be of any utility for longer AR models.

Here we show what one can do in those cases (this will be useful in many contexts so let's work it out in detail).

Let us study an $AR(2)$

$$x_t=\phi_1 x_{t-1}+\phi_2 x_{t-2}+\epsilon_t$$

Rewrite it in *companion form*:

$$\begin{bmatrix} x_t \\ x_{t-1}\end{bmatrix}=\begin{bmatrix}\phi_1 & \phi_2 \\ 1 & 0\end{bmatrix}\begin{bmatrix}x_{t-1}\\ x_{t-2}\end{bmatrix}+\begin{bmatrix}\epsilon_t\\ 0\end{bmatrix}$$

or, more compactly

$$X_t=AX_{t-1}+\varepsilon_t$$

This is a system of difference equations. How do we solve it?

A basic result from linear algebra says that a square matrix (like A) can be written in spectral form, $A=B\Lambda B^{-1}$, where the columns of B are the eigenvectors, and Λ is a diagonal matrix, with eigenvalues on the diagonal:

$$\Lambda=\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}$$

The system becomes:

$$X_t = B\Lambda B^{-1} X_{t-1} + \varepsilon_t$$

Premultiply the system by B^{-1}, to obtain

$$B^{-1}X_t = B^{-1}B\Lambda B^{-1}X_{t-1} + B^{-1}\varepsilon_t$$

Define $\tilde{X}_t = B^{-1}X_t$ and $\tilde{\varepsilon}_t = B^{-1}\varepsilon_t$
The system has become:

$$\tilde{X}_t = \Lambda \tilde{X}_{t-1} + \tilde{\varepsilon}_t$$

The interesting feature now is that the two elements of the vector $\tilde{X}_t$ are like two AR(1) (as Λ is diagonal) with a different forcing term.

We know how to solve them backwards, provided that $|\lambda_1| < 1$ and $|\lambda_2| < 1$: $\tilde{X}_{1t} = \sum_{j=0}^{\infty} \lambda_1^j \tilde{\varepsilon}_{1,t-j}$ and $\tilde{X}_{2t} = \sum_{j=0}^{\infty} \lambda_2^j \tilde{\varepsilon}_{2,t-j}$ or in matrix form:

$$\begin{bmatrix} \tilde{X}_{1t} \\ \tilde{X}_{2t} \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{\infty} \lambda_1^j \tilde{\varepsilon}_{1,t-j} \\ \sum_{j=0}^{\infty} \lambda_2^j \tilde{\varepsilon}_{2,t-j} \end{bmatrix}$$

We can recover the solution for the original variables from: $X_t = B\tilde{X}_t$. The first of the two elements of the vector X_t is x_t : here it is a solution to the $AR(2)$ difference equation.

It is clear that the conditions for the existence of a backwards solution are $|\lambda_1| < 1$ and $|\lambda_2| < 1$.

Note that the stability properties can also be obtained by solving the *characteristic equation*:

$$1 - \phi_1 z - \phi_2 z^2 = 0$$

and checking that the absolute value of the roots z_1 and z_2 is larger than 1.

It is important to be careful: if one uses the characteristic equation above, the condition for existence of a backwards solution is that the *the roots* z_1 *and* z_2 *are larger than one in absolute value.* If one uses the condition on the eigenvalues of A, the condition is that *the eigenvalues* λ_1 *and* λ_2 *are less than one in absolute value.*

The eigenvalues are those λ that solve the system of equations: $(A - \lambda I)b = 0$. This system of equations has a non trivial solution for b (the eigenvectors...) if and only if the matrix $(A - \lambda I)$ is non-invertible, that is if its determinant is equal to zero. The condition on the determinant of $(A - \lambda I)$ delivers the equation: $\lambda^2 - \phi_1 \lambda - \phi_2 = 0$). Note that the solutions of $1 - \phi_1 z - \phi_2 z^2 = 0$, call them z_1, z_2 are precisely $z_1 = 1/\lambda_1$ and $z_2 = 1/\lambda_2$

Obtaining the coefficients of the MA(∞) from an ARMA$(\mathbf{p}, \mathbf{q})$.

From:

$$\phi(L)x_t = \theta(L)\epsilon_t$$

$$x_t = \frac{\theta(L)}{\phi(L)}\epsilon_t = \psi(L)\epsilon_t$$

to find the elements in the $\psi(L)$ polynomial, one can equate the powers of L in:

$$\theta(L) = \psi(L)\phi(L)$$

Expectations and autocovariance functions for ARMA processes

Covariance-stationarity requires the existence and the time-independence of the expectation and of the autocovariance function. Any covariance-stationary stochastic process is uniquely characterized by an autocovariance function. This means that the autocovariance function is a very important tool. Let us compute expectations and autocovariance functions for some ARMA process.

MA(1): $x_t = \epsilon_t + \theta_1\epsilon_{t-1}$

The unconditional expectation:

$$E(x_t) = E(\epsilon_t + \theta_1\epsilon_{t-1}) = 0$$

The unconditional variance, γ_0:

$$\gamma_0 = Cov(x_t, x_t) = Cov(\epsilon_t + \theta_1\epsilon_{t-1}, \epsilon_t + \theta_1\epsilon_{t-1}) = \sigma_\epsilon^2 + \theta_1^2\sigma_\epsilon^2$$

$$\gamma_1 = Cov(x_t, x_{t-1}) = Cov(\epsilon_t + \theta_1\epsilon_{t-1}, \epsilon_{t-1} + \theta_1\epsilon_{t-2}) = \theta_1\sigma_\epsilon^2$$

$$\gamma_2 = Cov(x_t, x_{t-2}) = Cov(\epsilon_t + \theta_1\epsilon_{t-1}, \epsilon_{t-2} + \theta_1\epsilon_{t-3}) = 0$$

$$\gamma_k = Cov(x_t, x_{t-k}) = Cov(\epsilon_t + \theta_1\epsilon_{t-1}, \epsilon_{t-k} + \theta_1\epsilon_{t-k-1}) = 0, \forall k > 2$$

The conditional mean:

$$E(x_t|I_{t-1}) = \theta_1\epsilon_{t-1}$$

The conditional variance, $V(x_t|I_{t-1})$

$$\begin{aligned} V(x_t|I_{t-1}) &= E(x_t^2|I_{t-1}) - E(x_t|I_{t-1})^2 \\ &= E(\epsilon_t^2 + \theta_1^2\epsilon_{t-1}^2 + 2\theta_1\epsilon_t\epsilon_{t-1}|I_{t-1}) - \theta_1^2\epsilon_{t-1}^2 \\ &= \sigma^2 + \theta_1^2\epsilon_{t-1}^2 + 0 - \theta_1^2\epsilon_{t-1}^2 = \sigma^2 \end{aligned}$$

MA(2): $x_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2}$

The unconditional expectation:

$$E(x_t) = E(\epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2}) = 0$$

The unconditional autocovariance:

$$\gamma_0 = Cov(x_t, x_t) = Cov(\epsilon_t+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}, \epsilon_t+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}) = \sigma_\epsilon^2+\theta_1^2\sigma_\epsilon^2+\theta_2^2\sigma_\epsilon^2$$

$$\gamma_1 = Cov(x_t, x_{t-1}) = Cov(\epsilon_t+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}, \epsilon_{t-1}+\theta_1\epsilon_{t-2}+\theta_2\epsilon_{t-3}) = \theta_1\sigma_\epsilon^2+\theta_2\theta_1\sigma_\epsilon^2$$

$$\gamma_2 = Cov(x_t, x_{t-2}) = Cov(\epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2}, \epsilon_{t-2} + \theta_1\epsilon_{t-3} + \theta_2\epsilon_{t-4}) = \theta_2\sigma_\epsilon^2$$

$$\gamma_k = Cov(x_t, x_{t-k}) = Cov(\epsilon_t+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}, \epsilon_{t-k}+\theta_1\epsilon_{t-k-1}+\theta_2\epsilon_{t-k-2}) = 0, \forall k > 3$$

The autocovariance function of a $MA(p)$ is truncated after p terms.

The conditional mean:

$$E(x_t|I_{t-1}) = \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2}$$

The conditional variance, $V(x_t|I_{t-1})$:

$$V(x_t|I_{t-1}) = \sigma^2$$

MA(∞): $x_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ...$

The expectation is zero.
The variance, γ_0:

$$\gamma_0 = Cov(x_t, x_t) = Cov(\epsilon_t+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}+..., \epsilon_t+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}+...) = \sigma_\epsilon^2(1+\theta_1^2+\theta_2^2+\theta_3^2+...)$$

The autocovariance function of an MA(∞) is never truncated.

In order for this process to have a well defined variance, the condition of *square summability* of the coefficients must hold:

$$\sum_{j=0}^{\infty} \theta_j^2 < \infty$$

Notice that in the $MA(\infty)$ above I have assumed $\theta_0 = 1$.

Let us move to AR processes.

AR(1) with $|\phi_1| < 1 : x_t = \phi_1 x_{t-1} + \epsilon_t$

$$E(x_t) = E((1-\phi_1 L)^{-1}\epsilon_t) = 0$$

$$\gamma_0 = Cov(x_t, x_t) = Cov(\sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}, \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}) = \frac{\sigma_\epsilon^2}{1-\phi_1^2}$$

or, by exploiting stationarity: $V(x_t) = V(\phi_1 x_{t-1}) + V(\epsilon_t)$ from which: $V(x_t) = \phi_1^2 V(x_t) + \sigma_\epsilon^2$

Therefore:

$$\gamma_0 = \frac{\sigma_\epsilon^2}{1-\phi_1^2}$$

$$\begin{aligned}\gamma_1 &= Cov(x_t, x_{t-1}) = Cov(\epsilon_t + \phi_1\epsilon_{t-1} + \phi_1^2\epsilon_{t-2} + ..., \epsilon_{t-1} + \phi_1\epsilon_{t-2} + \phi_1^2\epsilon_{t-3} + ...) \\ &= \sigma_\epsilon^2\phi_1(1 + \phi_1^2 + \phi_1^4 + ...) = \frac{\phi_1}{1-\phi_1^2}\sigma_\epsilon^2 = \phi_1\gamma_0\end{aligned}$$

$$\gamma_2 = Cov(x_t, x_{t-2}) = \phi_1^2\gamma_0$$

$$\vdots$$

$$\gamma_k = Cov(x_t, x_{t-k}) = \phi_1^k\gamma_0$$

The autocovariance function of an AR(1) is a *geometrically* declining sequence.

The conditional mean, $E(x_t|I_{t-1})$:

$$E(x_t|I_{t-1}) = \phi_1 x_{t-1}$$

The conditional variance, $V(x_t|I_{t-1})$:

$$\begin{aligned}V(x_t|I_{t-1}) &= E(x_t^2|I_{t-1}) - E(x_t|I_{t-1})^2 \\ &= \phi_1^2 x_{t-1}^2 + \sigma^2 - \phi_1^2 x_{t-1}^2 = \sigma^2\end{aligned}$$

AR(2) with roots of $1 - \phi_1 z - \phi_2 z^2 = 0$ larger than one

For an AR(2) process, the above method becomes cumbersome. We can use the so-called *Yule-Walker recursions*

From

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t$$

Multiply both sides by x_t and take expectations:

$$E(x_t x_t) = \gamma_0 = \phi_1 E(x_t x_{t-1}) + \phi_2 E(x_t x_{t-2}) + E(x_t \epsilon_t) = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma_\epsilon^2$$

Multiply both sides by x_{t-1} and take expectations:

$$E(x_{t-1} x_t) = \gamma_1 = \phi_1 E(x_{t-1} x_{t-1}) + \phi_2 E(x_{t-1} x_{t-2}) + E(x_{t-1} \epsilon_t) = \phi_1 \gamma_0 + \phi_2 \gamma_1$$

Multiply both sides by x_{t-2} and take expectations:

$$E(x_{t-2} x_t) = \gamma_2 = \phi_1 E(x_{t-2} x_{t-1}) + \phi_2 E(x_{t-2} x_{t-2}) + E(x_{t-2} \epsilon_t) = \phi_1 \gamma_1 + \phi_2 \gamma_0$$

Multiply both sides by x_{t-3} and take expectations:

$$E(x_{t-3} x_t) = \gamma_3 = \phi_1 E(x_{t-3} x_{t-1}) + \phi_2 E(x_{t-3} x_{t-2}) + E(x_{t-3} \epsilon_t) = \phi_1 \gamma_2 + \phi_2 \gamma_1$$

and so on...

Collecting the resulting equations:

$$\begin{aligned} \gamma_0 &= \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma_\epsilon^2 \\ \gamma_1 &= \phi_1 \gamma_0 + \phi_2 \gamma_1 \\ \gamma_2 &= \phi_1 \gamma_1 + \phi_2 \gamma_0 \end{aligned}$$

The first 3 equations can be solved for γ_0, γ_1 and γ_2.
The higher autocovariances will follow the difference equation:

$$\gamma_{k+1} = \phi_1 \gamma_k + \phi_2 \gamma_{k-1}$$

If the roots are larger than one, this difference equation is stable and $\lim_{k \to \infty} \gamma_k = 0$

The conditional mean, $E(x_t | I_{t-1})$:

$$E(x_t | I_{t-1}) = \phi_1 x_{t-1} + \phi_2 x_{t-2}$$

The conditional variance, $V(x_t | I_{t-1})$:

$$V(x_t | I_{t-1}) = E(x_t^2 | I_{t-1}) - E(x_t | I_{t-1})^2 = \sigma^2$$

Causality and Invertibility

An ARMA process is *causal* if it can be expressed as:

$$x_t = \psi(L) \epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

where the past ϵ_{t-j} "cause" x_t

An ARMA process is *invertible* if it can be expressed as:

$$\epsilon_t = \eta(L)x_t = \sum_{j=0}^{\infty} \eta_j x_{t-j}$$

where ϵ_t can be expressed as function of past x_{t-j}

What is OLS estimating?
The least squares limit is $\frac{\gamma_1}{\gamma_0}$
If the true model is an AR(1), then the OLS estimate of an AR(1) is:

$$\frac{\frac{\phi\sigma^2}{1-\phi^2}}{\frac{\sigma^2}{1-\phi^2}} = \phi$$

If the true model is an AR(1) and we estimate an AR(2), $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$, OLS will estimate:

$$\begin{aligned}
\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} &= \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \\
&= \frac{1}{\gamma_0^2 - \gamma_1^2} \begin{bmatrix} \gamma_0 & -\gamma_1 \\ -\gamma_1 & \gamma_0 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \\
&= \frac{1}{\gamma_0^2 - \gamma_1^2} \begin{bmatrix} \gamma_0\gamma_1 - \gamma_1\gamma_2 \\ \gamma_0\gamma_2 - \gamma_1^2 \end{bmatrix} \\
&= \begin{bmatrix} \phi_1 \\ 0 \end{bmatrix}
\end{aligned}$$

The partial autocorrelation function

Suppose that we recursively compute the following regressions

$$x_t = \pi_1 x_{t-1} + \nu_{1t}$$

and save π_1. Then

$$x_t = \alpha_1 x_{t-1} + \pi_2 x_{t-2} + \nu_{2t}$$

and save π_2 and so on...
What do we learn by looking at the sequence $\{\pi_j\}_{j=1}^N$?

For an $AR(1) : x_t = \phi_1 x_{t-1} + \epsilon_t$

$\pi_1 = \phi_1$
$\pi_k = 0$ for $k > 1$

For an $AR(2)$: $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t$

$\pi_1 \neq 0$
$\pi_2 = \phi_2$
$\pi_k = 0$ for $k > 2$

and so on...

The coefficients of the partial autocorrelation function for an $AR(p)$ are zero after the p-th term.

For an $MA(1)$: $x_t = \epsilon_t + \theta_1 \epsilon_{t-1}$ we can substitute recursively:

$$x_t = \epsilon_t + \theta_1(x_{t-1} - \theta_1 \epsilon_{t-2}) = \epsilon_t + \theta_1[x_{t-1} - \theta_1(x_{t-2} - \theta_1 \epsilon_{t-3})] = \theta_1 x_{t-1} - \theta_1^2 x_{t-2} + ... + \epsilon_t$$

and so on...

All the coefficients of the partial autocorrelation function for an $MA(1)$ are non-zero and decay geometrically.

Let us collect what we have learnt:

- pure $AR(p)$: geometrically decaying autocorrelation function after the $p-th$ term. Truncated partial autocorrelation function at the $p - th$ term.
- pure $MA(q)$: truncated autocorrelation function after q terms. Geometrically decaying partial autocorrelation function after the $q - th$ term.

The Box-Jenkins approach

The objective of Box-Jenkins was to find a *good forecasting model* for the stationary time series x_t among $ARMA(p, q)$ (actually their approach is fine with non-stationary data as well, but we will see this later on).

The methodology is composed by the following steps:

- Identification (or, better, specification)
- Estimation
- Diagnostic checks (go back to specification if necessary)
- Forecast

Specification: this step boils down to the choice of the parameters p, q in an $ARMA(p, q)$.

Estimate the sequences $\{\gamma_s\}$ and $\{\pi_s\}$ and compare them to the typical shapes of basic models.

Remark: We said that the autocovariance function is a fundamental object for stationary stochastic processes: it identifies exactly the process. On the contrary the ARMA representation does not. This means that we may end up with two different ARMA models (with different p and q) both equally good.

As the Box-Jenkins methodology is designed to find a good forecasting model, it turns out that it is often better to use *parsimony* in the choice of

p and q. Large values for p and q imply many parameters to estimate. This will result in estimation uncertainty and will be transmitted to the forecasts. Sometimes in applied work it turns out that it is better to use the wrong values of p and q (smaller values than the "true" ones) in order to minimize the estimation error and obtain better forecasts.

2 problems:

1. it is not always easy to associate the shapes of γ_s and π_s to values for p and q.

2. the estimated $\hat{\gamma}_s$ and $\hat{\pi}_s$ may differ substantially from their theoretical counterparts (consistency is an asymptotic result).

There are tests for testing whether sample autocorrelations are statistically different from zero:

Under the null hypothesis that x_t follows an $MA(s-1)$, that is under the null that $\rho_s = 0$ from s on:

$$\hat{\rho}_s \rightarrow N(0, T^{-1}(1 + 2\sum_{j=1}^{s-1} \rho_j^2))$$

For example, under the null hypothesis that x_t is a white noise (that is, all $\rho_s = 0$, $s > 0$):

$$\begin{aligned} \hat{\rho}_1 &\rightarrow N(0, T^{-1}) \\ \hat{\rho}_2 &\rightarrow N(0, T^{-1}) \\ &\vdots \\ \hat{\rho}_s &\rightarrow N(0, T^{-1}) \end{aligned}$$

Under the null hypothesis that x_t is an $MA(1)$ (that is $\rho_s = 0$, $s > 1$):

$$\begin{aligned} \hat{\rho}_1 &\rightarrow N(0, T^{-1}(1 + 2\rho_1^2)) \\ \hat{\rho}_2 &\rightarrow N(0, T^{-1}(1 + 2\rho_1^2)) \\ &\vdots \\ \hat{\rho}_s &\rightarrow N(0, T^{-1}(1 + 2\rho_1^2)) \end{aligned}$$

The true value of ρ_1 which appear in the asymptotic variance will be replaced by the estimated value: $\hat{\rho}_1$.

If we want to test the null hypothesis: $H_0 : MA(0)$, the 95% confidence interval for ρ_1 is: $\hat{\rho}_1 \pm 1.96T^{-1/2}$. The confidence interval will be the same for all $\rho_s, s > 0$

If we want to test the null hypothesis $H_0 : MA(1)$, the 95% confidence interval for:ρ_2 is: $\hat{\rho}_2 \pm 1.96\sqrt{1 + 2\hat{\rho}_1^2}T^{-1/2}$. The confidence interval will be the same for all $\rho_s, s > 1$.

Other tests are based on the following test statistics:

Q-test (Box and Pierce):

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

Q-test (Ljung and Box):

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

Under the null hypothesis that all $\rho(k)$ from 1 to s are equal to zero, Q is asymptotically distributed as a χ^2_s.

Estimation

Suppose that after having computed $\hat{\gamma}_s$ and $\hat{\pi}_s$, and compared them to the typical shapes, we came out with values for p and q.

How do we estimate the parameters in $\phi(L)$ and $\theta(L)$?

For an $AR(p)$ model, no problem: the model satisfies the assumptions of OLS, so we can use standard OLS.

For an $MA(q)$ things are not that easy: we do not observe the variables on the right hand side! We have to use a different approach: maximum likelihood.

Maximum Likelihood

The likelihood of a set of random variables is their joint distribution: $f(X_1, X_2, ..., X_n, \theta)$. The joint distribution is assumed to be of known functional form (normal, χ^2, t, ecc.) and characterized by a vector of parameters θ (example: if the Xs are normally distributed, then $\theta = [\mu \quad \sigma^2]'$). The idea of maximum likelihood is to find the value of θ that maximizes the probability of having observed the data at hand (i.e. if the data come from a normal distribution with mean $\mu = 0$, the value of the density $f(x_1, x_2, ..., x_n, \mu = 100)$ will be lower than $f(x_1, x_2, ..., x_n, \mu = 0)$). Maximum likelihood requires an assumption on the distribution of the data (an assumption on the form of $f(.)$).

If the elements in the vector $X_1, X_2, ..., X_n$ are independent, the joint distribution can be written as the product of the marginals:

$$f(X_1, X_2, ..., X_n, \theta) = f(X_1, \theta) f(X_2, \theta) ... f(X_n, \theta)$$

If the elements in the vector $X_1, X_2, ..., X_n$ are *not independent*, the joint distribution can be written as the product of the conditional distributions and of the marginal distribution of X_1

$$f(X_1, X_2, ..., X_n, \theta) = f(X_1, \theta) f(X_2|X_1, \theta) f(X_3|X_1 X_2, \theta) ... f(X_n|X_1, ...X_{n-1}, \theta)$$

Let us work with a MA(1): $x_t = \theta\epsilon_{t-1} + \epsilon_t$ and assume that $\epsilon_t \sim iidN(0, \sigma^2)$

What does this assumption on ϵ_t imply for the distribution of the sample $x_1, x_2, ..., x_n$? As each x_t is a function of present and past ϵ_t, conditioning on x_{t-1} is equivalent to conditioning on ϵ_{t-1}. The conditional distribution of a generic x_t is:

$$f(x_t|x_{t-1}, \theta) = f(x_t|\epsilon_{t-1}, \theta) = N(\theta\epsilon_{t-1}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x_t - \theta\epsilon_{t-1})^2}{2\sigma^2}\right]$$

How do we compute ϵ_{t-1}? Let us start from ϵ_0 and assume that $\epsilon_0 = 0$[1] (this is the expected value for ϵ_0, right?).

From this, $f(x_1|\epsilon_0, \theta) = N(0, \sigma^2)$:

$$f(x_1|\epsilon_0, \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-{x_1}^2}{2\sigma^2}\right]$$

[1] This is the *conditional* likelihood. For more details, see Hamilton.

and $x_1 = \epsilon_1$. Given ϵ_1, we can compute: $x_2 - \theta\epsilon_1 = \epsilon_2$ whose distribution is:

$$f(x_2|\epsilon_1, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x_2 - \theta\epsilon_1)^2}{2\sigma^2}\right]$$

We can then compute: $x_3 - \theta\epsilon_2 = \epsilon_3$ and

$$f(x_3|\epsilon_2, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x_3 - \theta\epsilon_2)^2}{2\sigma_\epsilon^2}\right]$$

and so on...

The likelihood function of the sample $x_1, x_2, ..., x_n$ will therefore be:

$$f(x_1, x_2, ..., x_n, \theta) = f(x_1|\epsilon_0, \theta) f(x_2|\epsilon_1, \theta) .... f(x_n|\epsilon_{n-1}, \theta)$$

and for a given value of θ and σ^2 this is a number.

This function will be maximized with respect to the two unknown parameters, θ and σ_ϵ^2.

Warning: from the above iterations

$$\epsilon_t = y_t - \theta y_{t-1} + \theta^2 y_{t-2} - \theta^3 y_{t-3} + (-1)^t \theta^t \epsilon_0$$

the imposition of $\epsilon_0 = 0$ (and in general the conditional likelihood approach I describe here) will be innocuous as long as $|\theta| < 1$. In this case, θ^t will be small and the starting value will have no impact. If $|\theta| \simeq 1$ or $|\theta| > 1$, the starting value will be important and the results cannot be trusted.

Same logic, but more complicated for $ARMA(p, q)$. There will be standard errors associated with both the estimated $\hat{\phi}(L)$ and $\hat{\theta}(L)$.

Diagnostic Checks

After having estimated the model, one should check that the model "performs well".

1. The first thing to be tested is the white noise assumption of the residuals.

The idea is very easy: one should compute the estimated residuals, $\hat{\epsilon}_t$, compute their autocorrelation function and check that they do not display any remaining time dependence. If they do, they violate one assumption of the assumed $ARMA(p, q)$.

Suppose we estimate the autocorrelation function $\hat{\rho}(k)$ of the estimated residuals from an $ARMA(p, q)$. Under the null that all $\rho(k)$ from 1 to s are equal to zero, Q is now asymptotically distributed as a χ^2_{s-p-q}.

Q-test (Box and Pierce):

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

Q-test (Ljung and Box):

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

It can be shown that under the null hypothesis that the $\hat{\epsilon}_t$ follow an $MA(s-1)$, that is $\rho_s = 0$ from s on:

$$\begin{aligned} \hat{\rho}_1 &\rightarrow N(0, T^{-1}) \\ &\vdots \\ \hat{\rho}_s &\rightarrow N(0, T^{-1}(1 + 2\sum_{j=1}^{s-1} \rho_j^2)) \end{aligned}$$

If we want to test the null hypothesis that $\hat{\epsilon}_t$ is white noise ($H_0 : \rho_1 = 0$), the 95% confidence interval is: $\hat{\rho}_1 \pm 1.96T^{-1/2}$

If we want to test the null hypothesis $H_0 : \rho_2 = 0$, the 95% confidence interval is: $\hat{\rho}_2 \pm 1.96\sqrt{1 + 2\hat{\rho}_1^2}T^{-1/2}$

If we want to test the hypothesis that $\hat{\epsilon}_t$ follow a white noise (that is an $MA(0)$ basically), the above 95% confidence interval becomes:

$$\hat{\rho}_j \pm 1.96T^{-1/2}$$

for $j > 1$. This is the magnitude plotted by Eviews in the correlogram.

Jarque-Bera

LM Breusch-Godfrey

2. By adding more and more explanatory variables we will of course increase the in-sample fit (lower the sum of squared residuals), but, as said above, we may end up with an overparameterized model with a bad forecasting performance.

- Check fo the significance of the coefficients, by means of standard t-test or F-test.
- Reestimate the model, by eliminating those lags that are not significant.

3. There are statistics that take the opposite needs of fit and parsimony and trade them off: information criteria. On one side, the criteria reward fit, in the sense that the smaller the sum of squared residuals, the better. On the other side, the criteria penalize the estimation of too many parameters.

Here you have two of the many criteria available.

The criteria will suggest you to select the values for p and q so as to minimize the following magnitudes:

Akaike information criterion (AIC): $T\ln(\text{SSR}) + 2(1+p+q)$
Schwartz Bayesian criterion (SBC): $T\ln(\text{SSR}) + (1+p+q)\ln(T)$

The two effects are clearly discernible in the expressions. It is also clear that for $\ln(T) > 2$, the SBC will select smaller p and q than the AIC.

Forecasting

Suppose forecasting with an $AR(1)$: $x_t = \phi_1 x_{t-1} + \epsilon_t$
Shift the time index one period ahead, $x_{t+1} = \phi_1 x_t + \epsilon_{t+1}$
Conditional on the information available at time t, what is the forecast for x_{t+1}? And of x_{t+2}? Let us compute it...

$$\begin{aligned} E(x_{t+1}|I_t) &= \phi_1 E(x_t|I_t) + E(\epsilon_{t+1}|I_t) = \phi_1 x_t \\ E(x_{t+2}|I_t) &= \phi_1 E(x_{t+1}|I_t) + E(\epsilon_{t+2}|I_t) = \phi_1\phi_1 x_t = \phi_1^2 x_t \\ &\vdots \\ E(x_{t+k}|I_t) &= \phi_1^k x_t \end{aligned}$$

as $k \to \infty$, $E(x_{t+k}|I_t) \to 0$, the unconditional mean of x_t.

Suppose forecasting with an $AR(1)$ with a positive mean: $x_t = c + \phi_1 x_{t-1} + \epsilon_t$

$$\begin{aligned} E(x_{t+1}|I_t) &= c + \phi_1 x_t \\ E(x_{t+2}|I_t) &= c + \phi_1(c + \phi_1 x_t) = c(1+\phi_1) + \phi_1^2 x_t \\ &\vdots \\ E(x_{t+k}|I_t) &= c(1 + \phi + \phi^2 + ... + \phi^{k-1}) + \phi_1^k x_t \end{aligned}$$

as $k \to \infty$, $E(x_{t+k}|I_t) \to \dfrac{c}{1-\phi_1}$, the unconditional mean of x_t.

The forecast error is the error one does when forecasting x_{t+k} using information up to time t.
Let us study the properties of: $x_{t+k} - E(x_{t+k}|I_t)$
From:

$$x_{t+k} = \phi_1 x_{t+k-1} + \epsilon_{t+k} = \phi_1(\phi_1 x_{t+k-2} + \epsilon_{t+k-1}) + \epsilon_{t+k}$$

keeping on substituting, produces:

$$x_{t+k} = \phi_1^k x_t + \phi_1^{k-1}\epsilon_{t+1} + ... + \phi_1^2\epsilon_{t+k-2} + \phi_1\epsilon_{t+k-1} + \epsilon_{t+k}$$

We saw above that: $E(x_{t+k}|I_t) = \phi_1^k x_t$. The forecast error is

$$x_{t+k} - E(x_{t+k}|I_t) = \phi_1^{k-1}\epsilon_{t+1} + ... + \phi_1^2\epsilon_{t+k-2} + \phi_1\epsilon_{t+k-1} + \epsilon_{t+k}$$

The forecasts are unbiased

$$E(x_{t+k} - E(x_{t+k}|I_t)) = 0$$

The variance of the forecast error is increasing in k.

$$V(x_{t+k}-E(x_{t+k}|I_t)) = V(\phi_1^{k-1}\epsilon_{t+1}+...+\phi_1^2\epsilon_{t+k-2}+\phi_1\epsilon_{t+k-1}+\epsilon_{t+k}) = (\phi_1^{2(k-1)}+...+\phi_1^4+\phi_1^2+1)\sigma_\epsilon^2$$

As $k \to \infty$, $V(x_{t+k} - E(x_{t+k}|I_t)) \to \dfrac{\sigma_\epsilon^2}{1-\phi_1^2} = V(x_t)$, the unconditional variance of x_t.

If we assume that the ϵ_t are normally distributed, we can also compute the confidence intervals.

For the one step ahead forecast, $E(x_{t+1}|I_t) = c + \phi_1 x_t$, the forecast error is ϵ_{t+1}, and its variance is σ_ϵ^2. The confidence interval is then:

$$c + \phi_1 x_t \pm 1.96\sigma_\epsilon$$

For the two step ahead forecast, $E(x_{t+2}|I_t) = c(1+\phi_1) + \phi_1^2 x_t$, the forecast error is $\phi_1\epsilon_t + \epsilon_{t+1}$, with variance $(\phi_1^2+1)\sigma_\epsilon^2$

The confidence interval is then:

$$c(1+\phi_1) + \phi_1^2 x_t \pm 1.96(\phi_1^2+1)^{1/2}\sigma_\epsilon$$

Consider an ARMA(2,1): $x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t + \theta_1\epsilon_{t-1}$

The logic is the same as above. We simply use the fact that: $E_t(\epsilon_{t+k}) = 0, k > 0$

$$\begin{aligned}
x_{t+1|t} &= c + \phi_1 x_t + \phi_2 x_{t-1} + \theta_1\epsilon_t \\
x_{t+2|t} &= c + \phi_1 x_{t+1|t} + \phi_2 x_t \\
&\vdots \\
x_{t+k|t} &= c + \phi_1 x_{t+k-1|t} + \phi_2 x_{t+k-2|t}
\end{aligned}$$

How to evaluate the forecasting performance? Keep a part of the sample at the end and compute:

$$\begin{aligned}
RMSE &\quad \sqrt{\sum_{t=T+1}^{T+h} (\hat{y}_t - y_t)^2/h} \\
MAE &\quad \sum_{t=T+1}^{T+h} |\hat{y}_t - y_t|/h \\
MAPE &\quad \sum_{t=T+1}^{T+h} \frac{\hat{y}_t - y_t}{y_t}/h
\end{aligned}$$

Conditional vs unconditional forecast. It is always better to do conditional forecast than unconditional. Think of an AR(1): $y_t = c + \phi_1 y_{t-1} + \epsilon_t$. The one-step ahead unconditional forecast error is:

$$y_{t+1} - E(y_t) = y_{t+1} - \frac{c}{1-\phi_1}$$

The variance of the forecast error is:

$$E[(y_{t+1} - \frac{c}{1-\phi_1})^2] = (\sigma^2 + \phi_1^2\sigma^2 + \phi_1^4\sigma^2 + ...) = \frac{\sigma^2}{1-\phi_1^2}$$

The one step ahead conditional forecast error is:

$$y_{t+1} - E_t(y_{t+1}) = y_{t+1} - (c + \phi_1 y_t) = \epsilon_{t+1}$$

The variance of the forecast error is:

$$E[(y_{t+1} - c - \phi_1 y_t)^2] = E(\epsilon_{t+1}^2) = \sigma^2$$

It is clear that the variance of the conditional forecast error is smaller than the one of the unconditional.

Impulse response functions

Impulse response functions tell us what is the behaviour of $\{x_j\}$ when it is hit by a shock ϵ_t. They are useful tools when we want to understand what kind of dynamics is generated by the model.

Let us study an AR(1). Suppose the system in at rest. If a unitary shock hits the system at t, the paths for ϵ_t and x_t are:

time	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$
ϵ :	0	0	1	0	0	0
x :	0	0	1	ϕ_1	ϕ_1^2	ϕ_1^3

The system goes back to rest at speed given by ϕ_1

Let us study an MA(1). Suppose the system in at rest. If a unitary shock hits the system at t, the paths for ϵ_t and x_t are:

time	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$
ϵ :	0	0	1	0	0	0
x :	0	0	1	θ	0	0

Another way to think of impulse responses is to notice that the impulse response function is the same as

$$E_t(x_{t+k}) - E_{t-1}(x_{t+k})$$

It is the revision in the forecast for x_{t+k} caused by the new information arriving between $t-1$ and t:

$$\left.\begin{array}{c} E(x_t|x_{t-1}) = \phi x_{t-1} \\ E(x_t|x_t) = \phi x_{t-1} + \varepsilon_t \end{array}\right\} E(x_t|x_t) - E(x_t|x_{t-1}) = \varepsilon_t$$

$$\left.\begin{array}{c} E(x_{t+1}|x_{t-1}) = \phi^2 x_{t-1} \\ E(x_{t+1}|x_t) = \phi^2 x_{t-1} + \phi\varepsilon_t \end{array}\right\} E(x_{t+1}|x_t) - E(x_{t+1}|x_{t-1}) = \phi\varepsilon_t$$

and: $E(x_{t+k}|x_t) - E(x_{t+k}|x_{t-1}) = \phi^k \varepsilon_t$

Impulse responses for linear models (and therefore for ARMA) are symmetric: a positive and a negative shocks have the same effect with opposite sign.

The Wold Theorem

The Wold theorem states that: any zero-mean *covariance stationary* stochastic process $\{x_t\}_{t=-\infty}^{\infty}$ can be represented in the form:

$$x_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j} + \eta_t$$

where:

1. $\epsilon_t \equiv x_t - P(x_t|x_{t-1}, x_{t-2}, ...)$
2. $P(\epsilon_t|x_{t-1}, x_{t-2}..) = 0, E(\epsilon_t x_{t-j}) = 0,$
3. $E(\epsilon_t) = 0, E(\epsilon_t^2) = \sigma_\epsilon^2, E(\epsilon_t \epsilon_{t-j}) = 0$
4. The polynomial $\theta(L)$ is invertible
5. $\theta_0 = 1$ and $\sum_{j=0}^{\infty} \theta_j^2 < \infty$
6. $\{\theta_j\}$ and $\{\epsilon_j\}$ are unique
7. η_t is linearly deterministic: $\eta_t = P(\eta_t|x_{t-1}, x_{t-2}, ...)$

Remarks:

1. the ϵ_t are forecast errors (the symbol $P(a|b)$ denotes the fitted values of a linear regression of a on b)

2. lists the properties of forecast errors. They are orthogonal to the regressors

3. the ϵ_t are white noise

4. if $\theta(L)$ was not invertible, we could not express ϵ_t as a function of past and present x (see 1.)

5. this condition guarantees that x_t has finite variance

6. the Wold theorem delivers a *unique decomposition* of the process (as the autocovariances)

7. there may be parts of the process that are perfectly forecastable on the basis of past x_t.Remember the stationary but non-ergodic process we saw at the beginning? η_t is something of that kind...

Let us see the Wold decomposition for an AR(1) process

Suppose there is an information set at time 0: I_0 and that $P(x_1|I_0) = 0$
$\epsilon_1 \equiv x_1 - P(x_1|I_0) = x_1$
$\epsilon_2 \equiv x_2 - P(x_2|I_0, x_1)$
from which:

$$x_2 = P(x_2|I_0, x_1) + \epsilon_2 = \phi x_1 + \epsilon_2 = \phi\epsilon_1 + \epsilon_2$$

$\epsilon_3 \equiv x_3 - P(x_3|I_0, x_1, x_2)$
from which:

$$x_3 = P(x_3|I_0, x_1, x_2) + \epsilon_3 = \phi x_2 + \epsilon_2 = \phi(\phi\epsilon_1 + \epsilon_2) + \epsilon_3 = \phi^2\epsilon_1 + \phi\epsilon_2 + \epsilon_3$$

and so on...

Any stationary process has a Wold repr. This does not mean that the Wold representation is the true process.

The Wold says that we can express any stationary process as a linear combination of linear forecast errors ($P(a|b)$ expresses the fitted value of b as a linear function of a).The shocks in the Wold repr. need not be the true shocks.

This result is very important as any stationary process can be expressed as a linear combination of a white noise.

Any vector can be expressed as a linear combination of elements of a orthogonal basis:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = a/d \begin{bmatrix} d \\ 0 \\ 0 \end{bmatrix} + b/d \begin{bmatrix} 0 \\ d \\ 0 \end{bmatrix} + c/d \begin{bmatrix} 0 \\ 0 \\ d \end{bmatrix}$$

A stochastic process can be thought of an infinite vector of random variable. If the stochastic process is stationary, we can express it in the same form, in which the orthogonal basis is precisely constructed with white noise processes.

Non-stationarity

Economic data very often are non-stationary: either the first or the second moments are time-dependent.

If one looks at GDP over the last 50 years, it is obvious that the mean is not constant.

2 approaches to model trending series.

Example 1: the random walk

$$x_t = x_{t-1} + \varepsilon_t$$

Solve it backwards to a known value x_0 (you can also assume that $x_0 = 0$)

$$x_t = x_0 + \sum_{j=0}^{t} \varepsilon_j$$

The expectation $E(x_t) = x_0$
The variance: $V(x_t) = t\sigma_\varepsilon^2$
The random walk is not stationary
The autocorrelations: $\rho_s = \dfrac{Cov(x_t x_{t-s})}{\sqrt{V(x_t)}\sqrt{V(x_{t-s})}} = \dfrac{(t-s)\sigma_\varepsilon^2}{\sqrt{t\sigma_\varepsilon^2}\sqrt{(t-s)\sigma_\varepsilon^2}} = \dfrac{t-s}{\sqrt{t(t-s)}}$

As you see, the autocovariances vanish very slowly $\Rightarrow$ hard to disentangle a very persistent stationary AR from a random walk

The impulse response:

$$\lim_{k\to\infty} \frac{\partial x_{t+k}}{\partial \varepsilon_t} = 1$$

The effect of a shock never vanishes. The shock has a persistent effect on the series

The forecast

$$x_{t+k|t} = x_t$$

Example 2: the random walk with drift

$$x_t = \delta + x_{t-1} + \varepsilon_t$$

Solve it backwards starting from a known value x_0 (again, it could be zero)

$$x_t = x_0 + \delta t + \sum_{j=0}^{t} \varepsilon_j$$

The expectation $E(x_t) = x_0 + \delta t$
The variance: $V(x_t) = t\sigma_\varepsilon^2$

The autocorrelations: $\rho_s = \dfrac{t-s}{\sqrt{t(t-s)}}$

The random walk with drift is not stationary

The impulse response:

$$\lim_{k\to\infty} \frac{\partial x_{t+k}}{\partial \varepsilon_t} = 1$$

The effect of a shock does not vanish. The shock has a persistent effect on the series

The forecast:

$$x_{t+k|t} = x_t + \delta k$$

The forecast error:

$$x_{t+k} - x_{t+k|t} = (x_t + \delta k + \varepsilon_{t+k} + ... + \varepsilon_{t+1}) - (x_t + \delta k) = \varepsilon_{t+k} + ... + \varepsilon_{t+1}$$

with variance $k\sigma_\varepsilon^2$. The variance of the forecast error does not converge and keeps on increasing as the forecast horizon gets longer.

Example 3: a deterministic time-trend

$$x_t = \alpha + \beta t + \varepsilon_t$$

The expectation $E(x_t) = \alpha + \beta t$

The mean is not constant, but it is function of time.

The variance: $V(x_t) = \sigma_\varepsilon^2$

This series is non-stationary

The impulse response:

$$\lim_{k\to\infty} \frac{\partial x_{t+k}}{\partial \varepsilon_t} = 0$$

The forecast:

$$x_{t+k|t} = \alpha + \beta(t+k)$$

The forecast error:

$$x_{t+k} - x_{t+k|t} = \varepsilon_{t+k}$$

with variance σ_ε^2. The variance of the forecast error converges to the unconditional variance of the stationary part.

There is fundamental difference between cases 1 and 2 and case 3.

A shock in the deterministic time-trend model eventually vanishes and the series returns to the trend. A shock in models 1 and 2 never vanishes. The level is persistently modified by every shock.

The model in Case 3 displays a deterministic trend (we call it a trend-stationary (TS)). Models in Case 1 and 2 are called "stochastic trend" models.

They have a unit root in the AR part. They are also called $I(1)$ models (integrated of order one). They become stationary only after we take first differences and for this reason we call them difference-stationary (DS).

How to obtain stationarity in these two classes of models? One can follow 2 ways.

a. Subtract from the series a constant and a time trend

b. First-difference the series.

Case 1: $x_t = x_{t-1} + \varepsilon_t$

First difference:

$$\begin{aligned} x_t - x_{t-1} &= \varepsilon_t \\ y_t &= (1-L)x_t = \Delta x_t = \varepsilon_t \end{aligned}$$

The variable y_t is stationary. Good!

Case 2: $x_t = \delta + x_{t-1} + \varepsilon_t$

a. First difference:

$$\begin{aligned} x_t - x_{t-1} &= \delta + \varepsilon_t \\ y_t &= (1-L)x_t = \Delta x_t = \delta + \varepsilon_t \end{aligned}$$

Stationary. Good!

b. Subtract from x_t a constant and a linear trend. From $x_t = x_0 + \delta t + \sum_{j=0}^{t} \varepsilon_t$, it is clear that we will obtain:

$$y_t = x_t - x_0 - \delta t = \sum_{j=0}^{t} \varepsilon_t$$

The right hand side is still non-stationary! To be more precise, the mean is now constant, but the variance still depends on t.

$$\begin{aligned} E(y_t) &= 0 \\ V(y_t) &= t\sigma_\varepsilon^2 \end{aligned}$$

Case 3: a deterministic time-trend

$$x_t = \alpha + \beta t + \varepsilon_t$$

First difference x_t :

$$\begin{aligned}\Delta x_t &= \beta t - \beta t + \beta + \varepsilon_t - \varepsilon_{t-1} \\ &\beta + (1-L)\varepsilon_t\end{aligned}$$

The resulting series is stationary, but there is a non-invertible MA component.

If we subtract from x_t a constant and a time trend, the residuals will be stationary:

$$x_t - \alpha - \beta t = \varepsilon_t$$

We can also be more general, assuming the following model:

$$x_t = \alpha + \beta t + u_t$$

where $u_t \sim ARMA(p,q)$, that is: $\phi(L)u_t = \theta(L)\nu_t$ with an invertible MA polynomial $\theta(L)$.

Let us factor the AR polynomial as:

$$(1 - \phi_1 L - \phi_2 L^2 - ...) = (1 - \lambda_1 L)(1 - \lambda_2 L)...(1 - \lambda_p L)$$

If the roots $\lambda_1, \lambda_2, ...\lambda_p$ are all less than one in modulus, then $u_t = \frac{\theta(L)}{\phi(L)} = \psi(L)\nu_t$ and the model has the same properties of the trend stationary model (with a more complex transitory dynamics to converge to the trend).

If one root (say λ_1) is equal to one in modulus, then:

$$(1-L)u_t = \frac{\theta(L)}{(1-\lambda_2 L)...(1-\lambda_p L)}\nu_t = \frac{\theta(L)}{\phi^*(L)}\nu_t = \psi^*(L)\nu_t$$

and

$$(1-L)x_t = \beta + \psi^*(L)\nu_t$$

or

$$\phi^*(L)\Delta x_t = \chi + \theta(L)\nu_t$$

which is stationary. Hence the name "unit-root". Unit root processes have a unit root in the AR component and must be differentiated once in order to be stationary. A process like the one above is also called $ARIMA(p,d,q)$, where p and q are the lengths of the AR (not counting the unit roots) and MA polynomial respectively, while d indicates the number of unit roots in the AR polynomial (that is, the number of times x_t has to be differentiated to deliver a stationary $ARMA(p,q)$).

The underlying structure of the time series will be fundamentally different according to what is the form of the trend.

If we assume a TS model, then any shock will move the economy away from its trend only temporarily. If we assume a DS model, any shock will permanently shift the long run level.

Unit root tests

Given that TS and DS models have very different implications it is useful to develop statistical tests to distinguish between the two models.

There are interesting statistical issues in disentangling TS from DS models.

Case 1

Suppose the series is not trending but it is very persistent, with a positive sample mean. It could be so because it follows an AR(1) with a positive mean or because it has a unit root (and it happen that the particular realization has a positive mean).

A general rule for testing is that we should select a null and an alternative which are both "reasonable" characterization of the data at hand.

For example, a null hypothesis $H_0 : \phi = 1$ against an alternative $H_1 : \phi < 1$ in a regression $x_t = \phi x_{t-1} + \varepsilon_t$ is so and so... under the alternative, the series should display a zero mean.

A specification that can well characterize non trending data with a positive mean is:

$$x_t = \alpha + \phi x_{t-1} + \varepsilon_t \tag{1}$$

If $\phi < 1$, the series is stationary with a positive mean. When $\phi < 1$, we have a standard asymptotic result:

$$\sqrt{T}(\hat{\phi} - \phi) \rightarrow N(0, Avar(\hat{\phi}))$$

with $Avar(\hat{\phi}) = (1 - \phi^2)$.

If $\phi = 1$, the asymptotic distribution changes. The expression above does not converge to the standard normal distribution, but it has a non-standard non-symmetric (left-skewed) distribution (whose values have been tabulated by Dickey and Fuller, DF).

Therefore, when we want to test the null hypothesis $H_0 : \phi = 1$ against a one-sided $H_1 : \phi < 1$ using the t-ratio:

$$t = \frac{\hat{\phi} - 1}{\sqrt{\widehat{Var(\hat{\phi})}}}$$

we do not have to compare the value of the t-test to the standard normal distribution but to the non-standard DF distribution.

In the above case, the correct way to go is to test the joint hypothesis: $H_0 : \phi = 1$ and $\alpha = 0$. This F-test does not have a standard distribution and the correct p-values have been tabulated.

Why do we want to test also $\alpha = 0$? If $\phi = 1$ but $\alpha \neq 0$, x_t would be a unit root with drift. If we do not observe the drift we should restrict $\alpha = 0$.

In the first case ($x_t = \phi x_{t-1} + \varepsilon_t$), the model is misspecified under the alternative, in the second case ($x_t = \alpha + \phi x_{t-1} + \varepsilon_t$), we need to test for a joint restriction in order to have a null hypothesis that can characterize the observed behavior.

Remark: in EViews, the test is actually performed on the coefficients of the regression:

$$\Delta x_t = \alpha + \beta x_{t-1} + \varepsilon_t$$

where $\beta = \phi - 1$. This is obtained by subtracting x_{t-1} from both sides.

The null tested is $H_0 : \beta = 0$. There is no way to test directly the joint hypothesis $H_0 : \beta = 0$ and $\alpha = 0$. One has to estimate the above regression, compute the F statistic of the joint assumption and compare that value to the tables of the non-standard distribution tabulated by DF (in Hamilton, this is given at page 764).

Case 2

Suppose the series we observe is trending. The series can be trending either because of a deterministic time trend or because of a unit root with drift. Model (1) is not appropriate under the assumption of stationarity: a stationary AR(1) with positive mean is not trending.

Again, we should select a null and an alternative which are both "reasonable" characterization of the data at hand. A specification that nests both characterizations is:

$$x_t = \alpha + \beta t + \phi x_{t-1} + \varepsilon_t$$

Under the null hypothesis of a unit root $H_0 : \phi = 1$, if $\beta \neq 0$, the model above generates a quadratic trend:

$$x_t = x_0 + \alpha t + \beta t(t+1) + \sum_{j=0}^{t} \varepsilon_j$$

If $\phi = 1$ *and* $\beta = 0$, the model becomes a unit root with drift:

$$x_t = x_0 + \alpha t + \sum_{j=0}^{t} \varepsilon_j$$

Under the alternative hypothesis $H_1 : \phi < 1$ (with $\beta \neq 0$) the model above generates a linear trend (with $AR(1)$ transitory dynamics), the alternative model we are potentially observing.

The null of a unit root in this case should therefore be $H_0 : \phi = 1$ *and* $\beta = 0$.

The asymptotic distribution of the t-ratio is again non-standard and it is different from the non-standard distribution above.

In Eviews, the test is performed in the following regression:

$$\Delta x_t = \alpha + \beta t + \rho x_{t-1} + \varepsilon_t$$

where $\rho = \phi - 1$, in which $H_0 : \rho = 0$ *and* $\beta = 0$.

In general, the distribution of the test statistic under the null of a unit root is different according to the deterministic component (a constant or a time trend) assumed in the model and included in the test regression. See Hamilton or Enders for additional details.

A second important aspect of unit root tests is related to the presence of higher order autocorrelation in the residuals. Let us come back to the random walk against the AR(1). Not all the series can be modeled with AR(1) dynamics.

The series may be characterized as:

$$\phi(L)x_t = \alpha + \varepsilon_t$$

with: $1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p$ and we want to test if one of the roots is unity.

The above DF tests are not well suited in this situations, as the residuals will be autocorrelated.

We can write

$$\phi(L) = (1 - \rho L) - (\xi_1 L + \xi_2 L + ... + \xi_{p-1} L^{p-1})(1 - L)$$

where $\rho = \phi_1 + \phi_2 + ... + \phi_p$ and $\xi_i = -[\phi_{i+1} + ... + \phi_p]$, for $i = 1, 2, ...p - 1$[2]

From this:

$$\begin{aligned}(1 - \rho L)x_t - (\xi_1 L + \xi_1 L + ... + \xi_{p-1} L^{p-1})(1 - L)x_t &= \alpha + \varepsilon_t \\ (1 - \rho L)x_t &= \alpha + (\xi_1 L + \xi_1 L + ... + \xi_{p-1} L^{p-1})\Delta x_t + \varepsilon_t\end{aligned}$$

Finally,

$$x_t = \alpha + (\xi_1 L + \xi_1 L + ... + \xi_{p-1} L^{p-1})\Delta x_t + \rho x_{t-1} + \varepsilon_t$$

In the above regression, the test $H_0 : \rho = 1$ is performed. If $\rho = 1$ the series has a unit root.

[2] We will encounter below this decompositon: it is the Beveridge-Nelson decomposition. From $\phi(L)x_t = \varepsilon_t$, we can write: $x_t = A(L)x_{t-1} + \varepsilon_t$ and: $\Delta x_t = [A(L) - 1]x_{t-1} + \varepsilon_t$. Write $B(L) = [A(1) - 1] - [A(L) - 1] = A(1) - A(L)$. As $B(1) = 0$, it means that $B(L)$ has a unit root. Therefore, $B(L) = B^*(L)(1 - L)$ and $B^*(L)(1 - L) = A(1) - A(L)$. From this, $A(L) = A(1) - B^*(L)(1 - L)$ and: $\Delta x_t = [A(1) - 1]x_{t-1} - B^*(L)\Delta x_{t-1} + \varepsilon_t$. Note that $A(1) = \phi_1 + \phi_2 + ... + \phi_p = \rho$. As in the text, we could also move x_{t-1} to the right hand side: $x_t = A(1)x_{t-1} - B^*(L)\Delta x_{t-1} + \varepsilon_t$

The distribution of the F-test statistic in the test is non-standard and it is equal to the one in Case 1. The tests are called ADF (Augmented Dickey-Fuller).

Remark 1: in EViews, the test is actually performed on the regression:

$$\Delta x_t = \alpha + (\xi_1 L + \xi_1 L + ... + \xi_{p-1} L^{p-1})\Delta x_t + \beta x_{t-1} + \varepsilon_t$$

where $\beta = \rho - 1$. The null hypothesis of a unit root is then: $H_0 : \beta = 0$ and $\alpha = 0$..

Remark 2: Whether there are higher order terms or not, the distributions of the test statistics are unchanged. As above, the distributions depend to the deterministic components (constants and/or linear trends) in the models and in the assumed null hypothesis. One has to be careful!

Problem of UR tests: low power. They may not reject when the null is false

Testing the null of stationarity (the KPSS test)
Consider a model:

$$x_t = \alpha + \beta t + \phi x_{t-1} + \varepsilon_t$$

Fit the following regression:

$$x_t = \hat{\alpha} + \hat{\beta} t + \hat{\varepsilon}_t$$

Compute the partial sums:

$$S_\tau = \sum_{i=1}^{\tau} \hat{\varepsilon}_t$$

If the series is stationary, the variance of S_τ should be "small" (while it should be big under non-stationarity: $S_t = S_{t-1} + \varepsilon_t$). The test statistic is:

$$\eta = \frac{1}{T^2 s^2} \sum_{t=1}^{T} S_t^2$$

where s^2 is the long run variance (see Franses, p. 89). KPSS tabulated the distribution of the above statistic: if η is small, we cannot reject the null of stationarity. If η is big, we reject the null of stationarity against the altrernative of non-stationarity.

Spurious regression

From what we saw above it should be clear that when there is a unit root, standard asymptotic inference does not work.
Another manifestation of this is "spurious regression".

Suppose one runs a regression:

$$y_t = \alpha + \beta z_t + \nu_t \tag{2}$$

where:

$$\begin{aligned} y_t &= y_{t-1} + \varepsilon_{yt} \\ z_t &= z_{t-1} + \varepsilon_{zt} \end{aligned}$$

The two series are therefore completely unrelated (β should be equal to zero). It turns out that the standard t-test based on normality will reject $H_0 : \beta = 0$ in an extremely large number of cases and that the R^2 of the regression is very high. The problem is that the residuals from equation (2) are non-stationary. Note that we can write them as (suppose $\alpha = 0$):

$$\nu_t = y_t - \beta z_t = \sum_{j=1}^{t} \varepsilon_{yj} - \beta \sum_{j=1}^{t} \varepsilon_{zj}$$

the variance explodes as t increases. All the standard tests cannot be applied here. There will be a case in which the above regression will be meaningful, that is when ν_t is stationary. In that case, y_t and z_t will be said to be *cointegrated*.

Beveridge-Nelson decomposition

When time-series display a unit root, they display a permanent component (we have seen that shocks have permanent effect on the series).

It is therefore reasonable to devise procedures to separate one part that it is permanent, the non-stationary part, from a part that displays short run dynamics (stationary).

One way to do this is the Beveridge-Nelson decomposition.

Any unit root process, whose first differences can be written as an $MA(\infty)$ can be written as a sum of a random walk and a stationary process. Suppose y_t is a unit root process, whose Wold representation is:

$$\Delta y_t = \psi(L)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

It is possible to show that for any poliynomial $\psi(z)$ of order p there exists the following decomposition

$$\psi(z) = \psi(1) + \psi^*(z)(1 - z)$$

in which $\psi^*(z)$ is of order $p - 1$.

The polynomial $A(z) = \psi(z) - \psi(1)$ is still of order p. It is also true that $A(1) = 0$. This mean that 1 is a root of $A(z)$. This implies that $A(z) =$

$\psi^*(z)(1-z)$ in which $\psi^*(z)$ is of order $p-1$. In turn, $\psi^*(z)(1-z) = \psi(z)-\psi(1)$, so that:

$$\psi^*(z) = \frac{\psi(z) - \psi(1)}{1 - z}$$

One can show that:

$$\psi_j^* = \sum_{i=j+1}^{p} \psi_i$$

So:

$$\Delta y_t = \psi(1)\epsilon_t + \psi^*(L)(1 - L)\epsilon_t$$

Cumulating:

$$y_t = \underbrace{\psi(1) \sum_{j=1}^{t} \epsilon_j}_{permanent} + \underbrace{y_0 + \psi^*(L)\epsilon_t}_{transitory}$$

Note: y_t is a sum of two parts: a unit root part, function of ϵ_t and a stationary part. You can think of this as a trend-cycle decomposition, in which the trend is a random walk and the cycle is a process that moves around the unit root part.

Trend-cycle decompositions are not unique: there is an infinite number of ways in which one can separte a trend and a cycle (the concepts are not well-defined...). The Beveridge-Nelson decomposition assumes that both the trend and the cycle are driven by the same forces, the ϵ_t.

The magnitude $\psi(1)$ tells us the importance of the permanent component. If $\psi(1) = 0$, then $\psi(L) = \psi^*(L)(1 - L)$ and:

$$\Delta y_t = \psi^*(L)(1 - L)\epsilon_t$$

and:

$$y_t = \psi^*(L)\epsilon_t$$

The process has a $MA(\infty)$ also in levels. It is stationary also in levels!

Alternative derivation (from Hamilton). Suppose u_t is stationary:

$$u_t = \psi(L)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

Then:

$$u_1 + u_2 + ... + u_t = \psi(1)(\epsilon_1 + \epsilon_2 + ... + \epsilon_t) + \eta_t - \eta_0$$

where: $\psi(1) = \sum_{j=0}^{\infty} \psi_j, \eta_t = \sum_{j=0}^{\infty} \alpha_j \epsilon_{t-j}, \alpha_j = -(\psi_{j+1} + \psi_{j+2} + ...)$
If y_t is a unit root process, whose first difference is u_t then:

$$y_t = u_1 + u_2 + ... + u_t + y_0 = \psi(1)(\epsilon_1 + \epsilon_2 + ... + \epsilon_t) + \eta_t - \eta_0 + y_0$$

Trend-Cycle Decompositions
Linear trend
Hodrik-Prescott filter
Band-Pass filter

Modeling the conditional variance

Some economic time series are characterized by moments in which the variance changes.

A model that can capture this feature is the ARCH model, developed by Engle (1982):

$$\begin{aligned} \varepsilon_t &= \nu_t \sqrt{h_t} \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \end{aligned}$$

in which v_t is a white-noise process, with $E(\nu_t) = 0$ and $E(\nu_t^2) = 1$ and $\alpha_0 > 0$ and $0 < \alpha_1 < 1$.

Let us study the properties of $\{\varepsilon_t\}$.
Unconditionally, under stationarity:

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ E(\varepsilon_t^2) &= Var(\varepsilon_t) = E[\nu_t^2(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)] \\ Cov(\varepsilon_t, \varepsilon_{t-k}) &= 0 \end{aligned}$$

and

$$E(\varepsilon_t^2) = E(\nu_t^2)[\alpha_0 + \alpha_1 E(\varepsilon_{t-1}^2)]$$

from which: $E(\varepsilon_t^2) = \frac{\alpha_0}{1-\alpha_1}$. Notice: it is a variance and as such it cannot be negative! This is why we require $\alpha_0 > 0$ and $0 < \alpha_1 < 1$.
ε_t has the same properties of a white noise!

Conditionally,

$$\begin{aligned} E(\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, ...) &= E(\nu_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}) = 0 \\ E(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, ...) &= E[\nu_t^2(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2) | \varepsilon_{t-1}, \varepsilon_{t-2}, ...] = h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \end{aligned}$$

Note that as h_t is a conditional variance, it cannot be negative. For this reason, $\alpha_0 > 0$ and $\alpha_1 > 0$.

The conditional variance (which is the conditional expectation of ε_t^2,given that $E(\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, ...) = 0$) behaves like the conditional expectation of an AR(1) process.

The forecasts for ε_{t+k}^2, conditional on info up to time t:

$$\begin{aligned} E(\varepsilon_{t+1}^2 | \varepsilon_t) &= \alpha_0 + \alpha_1 \varepsilon_t^2 \\ E(\varepsilon_{t+2}^2 | \varepsilon_t) &= E(\nu_{t+2})[\alpha_0 + \alpha_1 E(\varepsilon_{t+1}^2 | \varepsilon_t)] = \alpha_0 + \alpha_1(\alpha_0 + \alpha_1 \varepsilon_t^2) = \alpha_0(1 + \alpha_1) + \alpha_1^2 \varepsilon_t^2 \\ E(\varepsilon_{t+k}^2 | \varepsilon_t) &= \alpha_0(1 + \alpha_1 + \alpha_1^2 + ... + \alpha_1^{k-1}) + \alpha_1^k \varepsilon_t^2 \\ \lim_{k \to \infty} E(\varepsilon_{t+k}^2 | \varepsilon_t) &= \frac{\alpha_0}{1 - \alpha_1} = \sigma_\varepsilon^2 \end{aligned}$$

In this model we have no correlation in level, but still correlation in the level squared. We can have clusters of large values (positive or negative) just because large values for ε_{t-1}^2 will be followed by large values for ε_t^2.

From

$$\varepsilon_t^2 = v_t^2(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)$$

The forecast error can be written as:

$$\eta_t = v_t^2 h_t - h_t = (v_t^2 - 1)h_t$$

From this we may write:

$$\varepsilon_t^2 = v_t^2(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2) = h_t + \eta_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \eta_t$$

which is an AR(1) process

Notice that the series ε_t can be the residual of an $ARMA$ model: $\varepsilon_t = \frac{\phi(L)}{\theta(L)} x_t$. This is an $ARMA(p,q) - ARCH(1)$ model

We combine a model for the conditional mean (standard $ARMA$) and models for the conditional variance ($ARCH$).

If the conditional mean is an $AR(1)$,

$$x_t = \phi x_{t-1} + \varepsilon_t$$

what are the properties of x_t?

$$\begin{aligned} E(x_t|x_{t-1}) &= \phi x_{t-1} \\ Var(x_t|x_{t-1}) &= Var(\varepsilon_t|x_{t-1}) = E(\varepsilon_t^2|x_{t-1}) = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \end{aligned}$$

The backwards solution is the usual one:

$$x_t = \frac{c}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}$$

The unconditional expectation is:

$$E(x_t) = \frac{c}{1-\phi_1}$$

The unconditional variance is:

$$V(x_t) = \frac{1}{1-\phi_1^2} \frac{\alpha_0}{1-\alpha_1}$$

$ARCH(p)$ process:

$$\begin{aligned} \varepsilon_t &= \nu_t \sqrt{h_t} \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + ... + \alpha_p \varepsilon_{t-p}^2 \end{aligned}$$

in which we restrict $\alpha_j > 0, \forall j$

$$E(\varepsilon_t^2 | I_{t-1}) = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + ... + \alpha_p \varepsilon_{t-p}^2$$

GARCH$(\mathbf{1},\mathbf{1})$

Often, the parameter p turns out to be large (parsimony...).

An alternative would be to model the "long" polynomial $\alpha(L)$ as arising from a ratio of two "short" polynomials: $\alpha(L)/\beta(L)$.

Let us now assume that:

$$\begin{aligned} \varepsilon_t &= \nu_t \sqrt{h_t} \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} \end{aligned}$$

in which v_t is a white-noise process, with $E(\nu_t) = 0$, $E(\nu_t^2) = 1$ and $\alpha_j > 0, \sum \beta_j > 0$ and $\alpha_j + \beta_j < 1$.

The conditional and unconditional mean of ε_t are zero (same steps as above).

The conditional variance of ε_t:

$$\begin{aligned} E(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, ...) &= E(\nu_t^2) E(h_t | \varepsilon_{t-1}, \varepsilon_{t-2}, ...) = h_t \\ E(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, ...) &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} \end{aligned}$$

Let us see in what sense we can interpret this model as an ARMA(1,1). From

$$\varepsilon_t^2 = v_t^2 (\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta h_{t-1})$$

The forecast error can be written as:

$$\eta_t = \varepsilon_t^2 - E(\varepsilon_t^2 | I_{t-1}) = v_t^2 h_t - h_t = (v_t^2 - 1) h_t$$

From this we write (adding and subtracting terms):

$$\begin{aligned} \varepsilon_t^2 &= v_t^2 h_t = v_t^2 h_t - h_t + (\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta h_{t-1}) + \beta \varepsilon_{t-1}^2 - \beta (v_{t-1}^2 h_{t-1}) \\ &= \alpha_0 + (\alpha_1 + \beta) \varepsilon_{t-1}^2 + \eta_t - \beta \eta_{t-1} \end{aligned}$$

which is an $ARMA(1,1)$ process in which the forcing term is a forecast error (it is a white noise).

Alternatively,

$$\begin{aligned}\eta_t &= \varepsilon_t^2 - h_t = \varepsilon_t^2 - (\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta h_{t-1}) + \beta \varepsilon_{t-1}^2 - \beta \varepsilon_{t-1}^2 \\ &= \varepsilon_t^2 - \alpha_0 - (\alpha_1 + \beta)\varepsilon_{t-1}^2 + \eta_t - \beta \eta_{t-1}\end{aligned}$$

If the conditional variance follows an $ARMA$, it should be possible to specify it by looking at autocovariances.

Once the conditional mean has been modeled with the standard $ARMA$ tools, we end up with white noise residuals.

The twist here is that the residuals squared should display $ARMA$ pattern and by looking at their ACF and $PACF$ we should be able to identify the orders of the lags in h_t and ε_t, using the same techniques developed above for the conditional mean (Bollerslev, 1986).

Testing for ARCH effects.

Model first the conditional mean and compute the residuals, $\hat{\varepsilon}_t$.
Estimate the regression:

$$\hat{\varepsilon}_t^2 = \zeta + \alpha_1 \hat{\varepsilon}_{t-1}^2 + \alpha_2 \hat{\varepsilon}_{t-2}^2 + ... + \alpha_m \hat{\varepsilon}_{t-m}^2 + e_t$$

and compute the centered R^2. It turns out that under the null that $\hat{\varepsilon}_t^2 \sim iidN(0, \sigma^2)$, the statistic $T \cdot R^2 \sim \chi_m^2$

(if the null is true, all the α's are equal to zero and the R^2 is zero). Alternatively, one can compute the F-test with $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$, distributed as an $F_{m,T-m}$.

An alternative is to implement an F-test in which the null is $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_m = 0$ and is distributed as a $F_{m,T-m}$.

The same idea can be used to test, say, an $ARCH(1)$ versus an $ARCH(4)$.
Estimate

$$\hat{\varepsilon}_t^2 = \zeta + \alpha_1 \hat{\varepsilon}_{t-1}^2 + \alpha_2 \hat{\varepsilon}_{t-2}^2 + \alpha_3 \hat{\varepsilon}_{t-3}^2 + \alpha_4 \hat{\varepsilon}_{t-4}^2 + e_t$$

and test $H_0 : \alpha_2 = \alpha_3 = \alpha_4 = 0$. This has an $F_{3,T-4}$ distribution.

Diagnostic checks

Goodness of the model for the mean: form the standardized residuals $\hat{s}_t = \frac{\hat{\varepsilon}_t}{\hat{h}_t^{1/2}}$ (squared residuals divided by their conditional variance). They should have zero mean and unitary variance. Test with Ljung-Box Q statistic. If you have a good model, $Q \simeq 0$.

Goodness of the model for the variance: form the standardized residuals squared $\hat{s}_t^2 = \frac{\hat{\varepsilon}_t^2}{\hat{h}_t} = v_t^2$ (residuals divided by their conditional variance). They should be white noise. Test with Ljung-Box Q statistic. Good model: $Q \simeq 0$.

Forecast for the levels

In general, the conditional mean will not depend on ARCH effect. ARCH effects will affect the confidence intervals

$$E(y_{t+1}|I_t) \pm 2\sqrt{h_{t+1}}$$

ARCH-in-mean

high risk $\Rightarrow$ high return. The conditional mean of returns should positively depend on risk (on variance)

$$\begin{aligned} y_t &= \delta h_t + \varepsilon_t \\ \varepsilon_t &= \nu_t \sqrt{h_t} \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \end{aligned}$$

δ captures the effect of higher variance of u_t on y_t.

Exponential GARCH

It is an asymmetric model

$$\begin{aligned} \varepsilon_t &= \nu_t \sqrt{h_t} \\ \log h_t &= \zeta + \sum_{j=1}^{\infty} \pi_j \left[|v_{t-j}| - E|v_{t-j}| + \varkappa v_{t-j} \right] \end{aligned}$$

The parameter $\varkappa$ captures a non-linear effect on h_t.

If $-1 < \varkappa < 0$ the effect of a positive shock is smaller than the effect of a negative shock.

If $\varkappa < -1$, a positive shock reduces volatility, a negative shock increases volatility.

$$\begin{aligned} \varepsilon_t &= \nu_t \sqrt{h_t} \\ \log h_t &= \zeta + \delta_1 \log h_{t-1} + ... + \delta_r h_{t-r} + \alpha_1 \left[|v_{t-1}| - E|v_{t-1}| + \varkappa v_{t-1} \right] + \\ &\quad ... + \alpha_m \left[|v_{t-m}| - E|v_{t-m}| + \varkappa v_{t-m} \right] \end{aligned}$$

Multivariate analysis

Many economic phenomena clearly deal with the joint behavior of more than one variables at a time. It is therefore important to develop models able to explain the joint behavior of more than one series. This is what we are going to do in this part of the course.

Let us start by defining few properties of a multivariate stochastic process. From now on, the stochastic process will be denoted as $\{x_t\}$, where x_t will be a $(n \times 1)$ vector of random variables.

The process $\{x_t\}$ will be weakly stationary if two condition hold:

$$\begin{aligned} E(x_t) &= \mu \\ E[(x_t - \mu)(x_{t-k} - \mu)'] &= \Gamma_k \end{aligned}$$

Γ_0 is the unconditional variance-covariance matrix of the vector x_t.
Γ_k is a covariance matrix and is a positive definite $(n \times n)$ matrix.
In a multivariate context: $\Gamma_k = \Gamma'_{-k}$ as, in general, $\Gamma_k \neq \Gamma_{-k}$.
Think of a bivariate case, $x_t = [x_{1,t} \;\; x_{2,t}]$, with $\mu = 0$:

$$\Gamma_1 = \begin{bmatrix} E(x_{1,t}x_{1,t-1}) & E(x_{1,t}x_{2,t-1}) \\ E(x_{2,t}x_{1,t-1}) & E(x_{2,t}x_{2,t-1}) \end{bmatrix}$$

and

$$\Gamma_{-1} = \begin{bmatrix} E(x_{1,t}x_{1,t+1}) & E(x_{1,t}x_{2,t+1}) \\ E(x_{2,t}x_{1,t+1}) & E(x_{2,t}x_{2,t+1}) \end{bmatrix}$$

where the element (1,2), $E(x_{1,t}x_{2,t-1})$ is equal to the element (2,1), $E(x_{2,t}x_{1,t+1})$ and the same for the other out-of-diagonal term.

The building block for multivariate processes is a vector white noise. ε_t is a vector white noise if:

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ E[\varepsilon_t \varepsilon_{t-k}{}'] &= \Sigma \text{ if } k = 0 \\ E[\varepsilon_t \varepsilon_{t-k}{}'] &= 0 \text{ if } k \neq 0 \end{aligned}$$

Note that there is no autocorrelation, nor cross-autocorrelation, but there may be contemporaneous correlation (Σ is not diagonal).

VAR processes

A VAR process is the multivariate counterpart of an AR.
Consider the expression:

$$x_t = c + A_1 x_{t-1} + A_2 x_{t-2} + ... + A_p x_{t-p} + \varepsilon_t$$

or more compactly,

$$A(L)x_t = c + \varepsilon_t$$

where: $A(L) = I - A_1 L - ... - A_p L^p$

This difference equation defines a $VAR(p)$. The lag operator in $A(L)$ has the same properties as in the scalar case.

We can therefore use the same logic as in the scalar case.

Let us work with a VAR(1): $x_t = c + Ax_{t-1} + \varepsilon_t$

Substituting backwards k times in the expression above,

$$x_t = A^{k+1}x_{t-k+1} + \varepsilon_t + A\varepsilon_{t-1} + A^2\varepsilon_{t-2} + ... + A^k\varepsilon_{t-k} + (I + A + A^2 + ... + A^k)c$$

As k grows, if the eigenvalues of A are all less than one in modulus, the first term disappears and the last term converges to

$$E(x_t) = \mu = c\sum_{j=0}^{\infty} A^j = (I - A)^{-1}c$$

the unconditional expectation of x_t.

The VAR(1) can therefore be written as a VMA(∞) representation:

$$x_t = \mu + \varepsilon_t + A\varepsilon_{t-1} + A^2\varepsilon_{t-2} + ... + A^k\varepsilon_{t-k} + ... = \mu + (I + AL + A^2L^2 + ... + A^kL^k + ...)\varepsilon_t$$

We can also write the VAR in deviations from the mean as:

$$x_t - \mu = A(x_{t-1} - \mu) + \varepsilon_t$$

We have seen in the univariate case that an equivalent way to state the condition of stationarity is that the roots of the characteristic equation are all larger than one in absolute value. The same is true here: the roots of $|I - Az| = 0$ must be larger than one to have stationarity.

As in the scalar case, when the eigenvalues are all less than one in absolute value, the VAR is stationary and has a solution in which x_t is function of past ε_ts.

If the condition for stationarity hold, we can "invert" the above AR polynomial to obtain the vector moving average representation (VMA): $x_t = \mu + A(L)^{-1}\varepsilon_t$.

The inversion of a matrix polynomial follows the same rules of the inversion of matrices.

If the VAR has more than one lag, checking the condition for stationarity and deriving the VMA representation is complicated. It is much easier to rewrite the $VAR(p)$ as a $VAR(1)$, deriving the companion form. Suppose x_t is of dimension $(n \times 1)$ and that it follows a stationary $VAR(3)$:

$$x_t = c + A_1x_{t-1} + A_2x_{t-2} + A_3x_{t-3} + \varepsilon_t$$

The companion form is:

$$\begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \end{bmatrix} = \begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} A_1 & A_2 & A_3 \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ x_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \end{bmatrix}$$
$$y_t = k + Cy_{t-1} + v_t$$

where y_t is of dimension $(3n \times 1)$ and C is of dimension $(3n \times 3n)$. The vector v_t is $(3n \times 1)$ but the rank of $E(v_t v_t')$ is only n.

We can study the stationarity of x_t by checking that all the eigenvalues of the matrix C lie within the unit circle.

The companion form is useful to compute the $VMA(\infty)$ representation of the VAR, as we know that the $VMA(\infty)$ for a $VAR(1)$, $x_t = Ax_{t-1} + \varepsilon_t$ is $x_t = C(L)\varepsilon_t$, where $C_i = A^i$.

The companion form will turn out to be extremely useful to compute impulse responses in a VAR

The companion form is also useful to compute the unconditional variance. Let us use the same example as above and assume $c = 0$.

$$\begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & A_3 \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ x_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \end{bmatrix}$$
$$y_t = Cy_{t-1} + v_t$$

If the VAR is stationary, one can multiply both sides by y_t' and take expectations to obtain:

$$\begin{aligned} E(y_t y_t') &= E(Cy_{t-1}y_{t-1}'C') + E(v_t v_t') \\ \Xi_0 &= C\Xi_0 C' + \Sigma \end{aligned}$$

which is a matrix equation in Γ_0. The solution to this equation is easy by using the Kronecker product and the *vec* operator.

Kronecker product: $\underset{(hg \times ks)}{C} = \underset{(h \times k)}{A} \otimes \underset{(g \times s)}{B} = \begin{bmatrix} a_{11}B & a_{12}B & ... & a_{1k}B \\ a_{21}B & & & \\ \vdots & & & \\ a_{h1}B & a_{h2}B & & a_{hk}B \end{bmatrix}$

vec operator: $vec \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} a & d & g & b & e & h & c & f & i \end{bmatrix}'$

Note that: $vec(ABC) = (C' \otimes A)vec(B)$

Let us apply the *vec* operator to the equation: $\Xi_0 = C\Xi_0 C' + \Sigma$

$vec(\Xi_0) = (C \otimes C)vec(\Xi_0) + vec(\Omega) = Fvec(\Xi_0) + vec(\Omega)$

This a a linear system of equations, whose solution is: $vec(\Xi_0) = (I - F)^{-1} vec(\Omega)$

(how do we know that $(I - F)$ is invertible?)

Note that:

$$\Xi_0 = E(\begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \end{bmatrix} \begin{bmatrix} x_t & x_{t-1} & x_{t-2} \end{bmatrix}) = \begin{bmatrix} \Gamma_0 & \Gamma_1 & \Gamma_2 \\ \Gamma_1' & \Gamma_0 & \Gamma_1 \\ \Gamma_2' & \Gamma_1' & \Gamma_0 \end{bmatrix}$$

From this, we can compute up to the $p - th$ autocovariance.

For the higher order autocovariances, note that:

$$\begin{aligned} E(y_t y_{t-1}') &= E(C y_{t-1} y_{t-1}') + E(v_t y_{t-1}') \\ \Xi_1 &= C\Xi_0 \end{aligned}$$

or in general

$$\begin{aligned} E(y_t y_{t-k}') &= E(C y_{t-1} y_{t-k}') + E(v_t y_{t-k}') \\ \Xi_k &= C\Xi_{k-1} = C^k \Xi_0 \end{aligned}$$

VARMA process

One may want to generalize the ARMA to vector values stochastic processes. While this is straightforward theoretically, there are difficult estimation problems wih VARMA models. Therefore, we don't say anything more on VARMA (keep in mind that in the last years there has been a resurgent interest in them ,as it often happens that the empirical representation of economic models is a VARMA. If that is the case, one has two ways to follow: either go the hard way and estimate the VARMA (with subspace algorithms, for example), or approximate the VARMA with a VAR in the hope that the approximation errror is not too big).

VAR Estimation

Estimating a VAR is extremely easy: you can do OLS equation by equation. A VAR actually is a SUR (Seemingly Unrelated Regressions) system. Think of a bivariate VAR(1)

$$\begin{aligned} x_t &= a_1 x_{t-1} + a_2 y_{t-1} + \varepsilon_{xt} \\ y_t &= b_1 x_{t-1} + b_2 y_{t-1} + \varepsilon_{yt} \end{aligned}$$

As one can see the only contemporaneous link between x_t and y_t is given by the correlation of the shocks. In the case in which the equations all have the same regressors the SUR estimator boils down to OLS equation by equation (Zellner).

Diagnostic checks are similar to those in the univariate case. Residuals have to be white noise. The choice of the lag length is done with information criteria.

What do you do with a VAR?

Forecasting

What is the best forecast for y_{t+1} given information up to time t?

$$y_{t+1|t} = \hat{y}_{t+1} = c + A_1 y_t + A_2 y_{t-1} + ... + A_p y_{t-p+1}$$

and K steps ahead?

$$y_{t+K|t} = \hat{y}_{t+K} = c + A_1 \hat{y}_{t+K-1} + A_2 \hat{y}_{t+K-2} + ... + A_p \hat{y}_{t+K-p}$$

The error in forecasting y_{t+1} given information up to time t is given by

$$y_{t+1} - y_{t+1|t} = \varepsilon_{t+1}$$

The variance of the one-step ahead forecast error is Ω.

The error in forecasting y_{t+2} given information up to time t is the difference

$$\begin{aligned} y_{t+2} - y_{t+2|t} &= (c + A_1 y_{t+1} + A_2 y_t + ... + A_p y_{t-p+1} + \varepsilon_{t+2}) - (c + A_1 \hat{y}_{t+1} + A_2 y_t + ... + A_p y_{t-p+1}) \\ &= A_1(y_{t+1} - \hat{y}_{t+1}) + \varepsilon_{t+2} = A_1 \varepsilon_{t+1} + \varepsilon_{t+2} \end{aligned}$$

The variance of the two-steps ahead forecast error is $A_1 \Omega A_1' + \Omega$

The error in forecasting y_{t+3} given information up to time t is the difference

$$\begin{aligned} & y_{t+3} - y_{t+3|t} \\ = \; & (c + A_1 y_{t+2} + A_2 y_{t+1} + ... + A_p y_{t-p+2} + \varepsilon_{t+3}) - (c + A_1 \hat{y}_{t+2} + A_2 \hat{y}_{t+1} + ... + A_p y_{t-p+2}) \\ = \; & A_1(y_{t+2} - \hat{y}_{t+2}) + A_2(y_{t+1} - \hat{y}_{t+1}) + \varepsilon_{t+3} \\ = \; & A_1(A_1 \varepsilon_{t+1} + \varepsilon_{t+2}) + A_2 \varepsilon_{t+1} + \varepsilon_{t+3} \\ = \; & (A_1^2 + A_2)\varepsilon_{t+1} + A_1 \varepsilon_{t+2} + \varepsilon_{t+3} \end{aligned}$$

and so on...

As in the univariate case, it is much easier to compute these magnitudes from the VMA representation). From $x_t = \Psi(L)\varepsilon_t$

$$x_{t+1} - x_{t+1|t} = \varepsilon_t$$

with variance Ω.

$$x_{t+2} - x_{t+2|t} = \varepsilon_{t+2} + \Psi_1 \varepsilon_{t+1} + \Psi_2 \varepsilon_t ... - (\Psi_2 \varepsilon_t + \Psi_3 \varepsilon_{t-1}) = \varepsilon_{t+2} + \Psi_1 \varepsilon_{t+1}$$

with variance $\Omega + \Psi_1 \Omega \Psi_1'$.

$$x_{t+3} - x_{t+3|t} = \varepsilon_{t+3} + \Psi_1\varepsilon_{t+2} + \Psi_2\varepsilon_{t+1} + \Psi_3\varepsilon_t ... - (\Psi_3\varepsilon_t + \Psi_4\varepsilon_{t-1}) = \varepsilon_{t+3} + \Psi_1\varepsilon_{t+2} + \Psi_2\varepsilon_{t+1}$$

with variance $\Omega + \Psi_1\Omega\Psi_1' + \Psi_2\Omega\Psi_2'$
and so on...

Check that the forecast errors computed from the VAR and the VMA are the same, for example, in a VAR(1).

Granger causality

The word "causality" evokes philosophical ideas... what causes what?
In time-series analysis, Granger has introduced a concept of "causality", in which the basic idea is that the cause has to happen before the effect in time (we will see that this definition sometimes creates problems).

Definition 1 *A variable w_t "Granger-cause" y_t if past values of w_t help in predicting y_t, given the past of y_t*

Consider a bivariate VAR

$$\begin{aligned} x_t &= a_1(L)x_{t-1} + a_2(L)y_{t-1} + \varepsilon_{xt} \\ y_t &= b_1(L)x_{t-1} + b_2(L)y_{t-1} + \varepsilon_{yt} \end{aligned}$$

In this example, x_t does not Granger cause y_t if $b_1(L) = 0$. One can therefore test for $b_1(L) = 0$ with an F-test (under stationarity).

Limits:
1. does have some funny implications. Weather forecast "causes" weather. Christmas cards "cause" Christmas, etc... The concept of causation used has to be kept in mind.

2. the results depend on the conditioning set, on the variables one is regressing on. One may find that a variable Granger cause another one in a bivariate system but may obtain a different result in a trivariate system.

Some economics

VAR were introduced in economics by Sims in his famous article "Macroeconomics and Reality", published in Econometrica in 1980[3].

[3] This is what I wrote about "Macroeconomics and Reality" in the first version of these notes in the Spring of 2009: "read it if you want to understand how you can win a Nobel prize... he will get one soon for that article....". Prof. Sims earned the Nobel Prize in Economics in 2011, toghether with Prof. Sargent.

Before that article a standard macroeconometric model was composed by a huge set of equations, specified one by one in which:

1. some variables were determined within the model and other variables were **assumed** to be exogenous.
2. some lags of some variables were **a priori excluded** from some equations.

A workable example is the following:

$$\begin{aligned} q_d &= \alpha_1 p + \alpha_2 y + \varepsilon_d \quad \text{demand equation} && (3)\\ q_s &= \beta_1 p + \varepsilon_s \quad \text{supply equation} && (4)\\ q_d &= q_s = q \quad \text{equilibrium} && (5) \end{aligned}$$

in which the error terms display the following properties:

$$E(\varepsilon_d) = E(\varepsilon_s) = 0, E(\varepsilon_d^2) = \sigma_d^2, E(\varepsilon_s^2) = \sigma_s^2 \text{ and } E(\varepsilon_s \varepsilon_d) = 0$$

The three equations above represent an economic "theory" (admittedly simple...) and they are called the *structural model*.

q and p are endogenous: their value is determined by the interaction of demand and supply.
y is exogenous: its value is determined outside the model.

The researcher is interested in the structural parameters, α_1 and β_1, the slopes of the demand and supply curves.
Suppose for a second that y is not there ($\alpha_2 = 0$). What happens if we run OLS on the above equations?

Consider the first equation, (in which I've used the equilibrium condition)

$$q_t = \alpha_1 p_t + \varepsilon_{dt}$$

In order for OLS to deliver consistent estimates, the predeterminedness conditions on the regressors must hold: $E(p_t \varepsilon_{dt}) = 0$.
The condition does not hold here.

Let us rewrite the system in **reduced form** (the endogenous as a function of exogenous and shocks), solving for q and p

$$\begin{aligned} p &= \frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1} = v_p \\ q &= \frac{\beta_1 \varepsilon_d - \alpha_1 \varepsilon_s}{\beta_1 - \alpha_1} = v_q \end{aligned}$$

From this, it is clear that

$$E(p_t\varepsilon_{dt}) = \frac{\sigma_d^2}{\beta_1 - \alpha_1} \neq 0$$

OLS is not consistent for α_1. It is also clear that

$$E(p_t\varepsilon_{st}) = \frac{\sigma_s^2}{\beta_1 - \alpha_1} \neq 0$$

OLS is not consistent for β_1.

Suppose one does not realize this and run OLS. What will he/she be estimating?

The OLS estimator is $Cov(p,q)/Var(p)$. The numerator will be:

$$Cov(\frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1}; \frac{\beta_1\varepsilon_d - \alpha_1\varepsilon_s}{\beta_1 - \alpha_1}) = \frac{\beta_1\sigma_d^2 + \alpha_1\sigma_s^2}{(\beta_1 - \alpha_1)^2}$$

and the regression coefficient:

$$\frac{Cov(p,q)}{Var(p)} = \frac{\frac{\beta_1\sigma_d^2 + \alpha_1\sigma_s^2}{(\beta_1 - \alpha_1)^2}}{\frac{\sigma_d^2 + \alpha_1^2}{(\beta_1 - \alpha_1)^2}} = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_s^2}\beta_1 + \frac{\sigma_s^2}{\sigma_d^2 + \sigma_s^2}\alpha_1$$

OLS will estimate a weighted average of the slope of supply and demand. This problem is called *simultaneous equation bias.*

The problem can be seen from a different angle by looking at the reduced form:

$$\begin{aligned} p &= \frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1} = v_p \\ q &= \frac{\beta_1\varepsilon_d - \alpha_1\varepsilon_s}{\beta_1 - \alpha_1} = v_q \end{aligned}$$

The statistical properties of the 2 variables p and q are characterized by 3 elements: the 2 variances and the covariance (see the reduced form). The demand-supply model requires the estimation of 4 parameters: α_1, β_1 and the variances of the 2 shocks, σ_s^2 and σ_d^2. In this case, there is no way to map the 3 reduced form coefficents in any of the 4 structural parameters. There is an *identification problem*: **the statistical features of the data are compatible with many structural models**. Note however that having the same number of reduced form and structural parameters is neither sufficient, nor necessary to obtain identification of the structural parameters.

How one can estimate the structural parameters in the various equations of such a system?

One way is to make the theory more stringent, for example assuming that the supply curve takes the form $\beta_1 p + \varepsilon_s = 0$ (I know this looks weird, but for the sake of the argument, forgive me). In this case, it is clear that the reduced form is:

$$\begin{aligned} p &= -\frac{\varepsilon_s}{\beta_1} \\ q &= -\frac{\alpha_1}{\beta_1}\varepsilon_s + \varepsilon_d \end{aligned}$$

It is clear here that $E(p\varepsilon_d) = 0$ and that α_1 can be estimated from:

$$q = \alpha_1 p + \varepsilon_d$$

Note here that we have 3 moments, 4 parameters to be estimated but only two (α_1 and σ_d^2) can be estimated.

A second way is to enrich the theory by adding variables that are uncorrelated with the error terms. Let us introduce y back in the model. y is what is called an *instrumental variable.* Let's assume $E(y_t\varepsilon_{dt}) = E(y_t\varepsilon_{st}) = 0$.

The reduced form becomes:

$$\begin{aligned} p &= \frac{\alpha_2}{\beta_1 - \alpha_1}y + \frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1} = \pi_1 y + v_p \\ q &= \frac{\beta_1\alpha_2}{\beta_1 - \alpha_1}y + \frac{\beta_1\varepsilon_d - \alpha_1\varepsilon_s}{\beta_1 - \alpha_1} = \pi_2 y + v_q \end{aligned}$$

One can estimate the two equations above with OLS (y is predetermined) and obtain $\hat{\pi}_1$ and $\hat{\pi}_2$. Note that the ratio $\dfrac{\hat{\pi}_2}{\hat{\pi}_1}$ is equal to β_1. We have an estimator for the slope of the supply curve. This is called *indirect least squares.*

We can also derive an *instrumental variable* (IV) estimator.

Recall that OLS are based on the so-called *normal equations,* derived from the predeterminedness conditions on the regressors:

$$\begin{aligned} E(x\varepsilon) &= 0 \\ E(x(y - x\beta)) &= 0 \\ E(xy - xx\beta)) &= 0 \\ E(xy) - E(xx)\beta &= 0 \\ \beta &= E(xx)^{-1}E(xy) \end{aligned}$$

In the supply curve, we can do the same using the instrument y :

$$\begin{aligned} E(y\varepsilon_s) &= 0 \\ E(y(q-\beta_1 p)) &= 0 \\ E(yq)-\beta_1 E(yp) &= 0 \\ \beta_1 &= E(yp)^{-1}E(yq) \end{aligned}$$

This is the instrumental variable (IV) estimator. If we do the same on the demand curve, we see that we cannot go far:

$$\begin{aligned} E(y\varepsilon_d) &= 0 \\ E(y(q-\alpha_1 p-\alpha_2 y)) &= 0 \\ E(yq)-\alpha_1 E(yp)-\alpha_2 E(yy) &= 0 \end{aligned}$$

??? one equation in 2 unknowns.... ???

In order to estimate the slope of the demand curve we would need an instrumental variable that shifts the supply and not the demand curve (try and see!).

We can also follow a third route. First, regress p on y and compute the fitted values $\hat{p} = E(yy)^{-1}E(yp)y$

Second, regress q on $\hat{p}$. This is: $E(\hat{p}\hat{p})^{-1}E(q\hat{p})$ (show that $E(\hat{p}\hat{p})^{-1}E(q\hat{p}) = E(yp)^{-1}E(qy)$) This estimator is called the two stage least squares (2SLS) estimator. The first step "cleans" the endogenous regressor p from the part that is not correlated with y and retains the part that is correlated with it. The second step simply regress q on that portion of p which is correlated with y.

There are close relations between these three approaches.

Note that if y entered also in the demand equation, it would be impossible to estimate the structural parameters of the supply curve (try!). We have used an *exclusion restriction*, which is the fact that by assumption y does not enter in the supply equation.

The same ideas can be approached using a matrix notation.

Suppose the structural form is:

$$Ay_t = \Gamma x_t + u_t$$

y are the endogenous, x are the predetermined (in x_t we can have lags of y_t).

The reduced form is

$$\begin{aligned} y_t &= A^{-1}\Gamma x_t + A^{-1}u_t \\ y_t &= \Pi x_t + v_t \end{aligned}$$

in which:

$$\begin{aligned} \Pi &= A^{-1}\Gamma \\ v_t &= A^{-1}u_t \quad \text{or} \quad Cov(v_t) = \Omega = A^{-1}\Sigma A^{-1\prime} \end{aligned}$$

The structural parameters of equation j are identified if given the couple $\{\Pi, \Omega\}$ one can recover uniquely A_j and Γ_j (the $j-th$ row of the matrices A and Γ). The conditions to recover the structural parameters for equation j are:

1. the parameters in equation j, in which there are d non-predetermined variables, can be estimated (and are therefore identified), if there are at least d instruments (order condition).
2. the instruments must contain non-redundant information (rank condition).

Identification in the model above was therefore achieved by contemporaneous exclusion restrictions and exogeneity assumptions.

Note that many different structural models can have the same reduced form. All the structural models with the same reduced form are indistinguishable on the basis of the data.

This was the state of the art up to the mid-70s. In the 1970s, 2 main things happened in economics

1. rational expectations $\Rightarrow$ equations are related and there are very few "zeros", especially on the lags
2. general equilibrium $\Rightarrow$ no more endogenous vs. exogenous. Everything is endogenous

This means that the two main assumptions on which models were built become all of a sudden obsolete and difficult to buy (Sims used a clearer expression: they were "incredible restrictions").

Structural VAR (SVAR) were according to Sims the response to all criticisms. In a SVAR all the variables are modelled as endogenous and there is no need to impose restrictions on the coefficients on past regressors. Economic theory entered only very mildly in a SVAR (Sims used the expression: "let the data speak").

A SVAR is a system of equations of the form (I assume only one lag and no constant for simplicity):

$$Ax_t = Bx_{t-1} + u_t \qquad u_t \sim VWN(0, \Sigma)$$

With reduced form:

$$x_t = Cx_{t-1} + v_t \qquad v_t \sim VWN(0, \Omega)$$

where $C = A^{-1}B$ and $v_t = A^{-1}u_t$ or in terms of variances: $\Omega = A^{-1}\Sigma A^{-1\prime}$.

As one can see, its structure is very similar to the structure of a simultaneous equation system.

As such we may think of following the same approach as above. Either we can assume that some of the elements in x_t are instruments (as in the example above, instruments often takes the form of exogenous variables and we do not like them) or we impose exclusion restrictions in the system. We do not like putting zeros on the lags (dynamics should be left unrestricted). The only place where we are allowed to put zeros is either on the contemporaneous relations between variables (the matrix A) or on the covariance between shocks (Σ).

How many reduced form parameters can we estimate from the data? How many structural parameters do we have in the structrual form? In the reduced form we have n^2 parameters in C and $n(n+1)/2$ distinct elements in Ω. In the structure we have n^2 in A, n^2 in B and $n(n+1)/2$ in Σ, in total, $2n^2+n(n+1)/2$. There are n^2 more parameters in the structural model than in the reduced form.

In order to hope to find a mapping between the reduced form and the structural form (recall indirect LS...), we need to impose n^2 restrictions on the structural form. Where? We said not on the dynamics (not on B). We will therefore restrict Σ and A.

From an economic point of view, it seems reasonable to assume that the sources of fluctuations in the structural system, the u_t are uncorrelated (demand shifter can reasonably be assumed to be orthogonal to supply shifter).

The assumption of orthogonal shocks gives some sense to the question "what happens to the variables in the model in response to a shock", say a technology shock or a monetary policy shock, or a fiscal shock. If the shocks were correlated, it would not be meaningful to study what happen when a positive supply shock arrives, as all the others move because of the correlation. A standard set of restrictions is therefore: $\Omega = I$ (the ones on the diagonal are a useful normalization).This amount to imposing $n(n+1)/2$ restrictions.

We had to impose n^2, we imposed $n(n+1)/2$. We still need at least $n(n-1)/2$ restrictions to hope to obtain a one-to-one mapping. These restrictions are often imposed on the matrix A.

Recall once again that any set of exactly identifying restrictions will deliver the same reduced form so from a statistical point of view, they are all the same.

The proposal of Sims was the following: assume that the matrix A is lower triangular. Note that this assumption amounts to the imposition of exactly $n(n-1)/2$ exclusion restrictions.

In a trivariate system this correspond to a structural model:

$$\begin{bmatrix} a_1 & 0 & 0 \\ b_1 & b_2 & 0 \\ c_1 & c_2 & c_3 \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{bmatrix} = B \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \\ x_{3t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix} \qquad u_t \sim VWN(0, I)$$

with reduced form:

$$\begin{bmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{bmatrix} = \begin{bmatrix} d_1 & 0 & 0 \\ e_1 & e_2 & 0 \\ f_1 & f_2 & f_3 \end{bmatrix} B \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \\ x_{3t-1} \end{bmatrix} + \begin{bmatrix} d_1 & 0 & 0 \\ e_1 & e_2 & 0 \\ f_1 & f_2 & f_3 \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix} \qquad v_t \sim VWN(0, \Omega)$$

$$\begin{bmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{bmatrix} = C \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \\ x_{3t-1} \end{bmatrix} + \begin{bmatrix} v_{1t} \\ v_{2t} \\ v_{3t} \end{bmatrix}$$

This identification scheme imposes a recursive structure to the system. The first shock, u_{1t} hits all the variables at t. The second shock, u_{2t} hits only the second and the third variable at t and will affect the third only with one period delay. The third shock affects only the third variable at t and will impact the first and the second with one period delay

By the same token, the contemporaneous relations between variables have a recursive structure. In the first equation there are no contemporaneous relations between the elements in x. In the second equation, there appear both x_1 and x_2 and in the third all the three.

How can we estimate this structural system? We can estimate the reduced form and solve the system of equations

$$\begin{aligned} C &= A^{-1}B \\ \Omega &= A^{-1}A^{-1\prime} \end{aligned}$$

This is a system of n^2 equations in n^2 unknowns and one can show that it has a unique solution.

It turns out that any positive definite matrix (such as the covariance matrix Ω) can be written as $\Omega = GG'$, where G is lower triangular and it is called the Choleski factor of Ω.

One can therefore decompose $\Omega = GG'$ and note that: $G = A^{-1}$ and that G^{-1} is a lower triangular matrix.

The structural form is therefore

$$G^{-1}x_t = G^{-1}Cx_{t-1} + G^{-1}v_t$$

in which there is a contemporaneous triangular structure and $Cov(u_t) = Cov(G^{-1}v_t) = G^{-1}\Omega G^{-1\prime}$, from which it is clear that $Cov(u_t) = I$.

Note that we can have an infinite number of structural models compatible with the same reduced form. The data will not tell me which one has a better fit, as all these models will be characterized by the same reduced form.

A standard example of this identification scheme is a VAR in which one shock is assumed to be a monetary policy shock.

The SVAR (the theoretical model) is:

$$\begin{bmatrix} a_1 & 0 & 0 \\ b_1 & b_2 & 0 \\ c_1 & c_2 & c_3 \end{bmatrix} \begin{bmatrix} Y_t \\ \pi_t \\ FFR_t \end{bmatrix} = B \begin{bmatrix} Y_{t-1} \\ \pi_{t-1} \\ FFR_{t-1} \end{bmatrix} + \begin{bmatrix} u_{Y,t} \\ u_{\pi,t} \\ u_{FFR,t} \end{bmatrix}$$

with $u_t \sim VWN(0, I)$. The third equation is assumed to be the reaction function of the Fed, the Fed sets the FFR at t, by looking at output and inflation at t. The first and second equation are not given a full structural interpretation and are thought of summarizing the dynamic relations between variables.

$u_{FFR,t}$ is the monetary policy shocks, that is, the non-systematic part of monetary policy, the part that is not related to the variables that are in the information set of the Fed. The triangularity assumption in the first two equations implies that Y_t and π_t do not respond at time t to a monetary policy shock at time t. The motivation for those exclusion restrictions is that it takes time for output and inflation to move in response to a variation in the FFR. One may rationalize it with some form of sticky prices in which prices move only slowly and output responds slowly as well.

An alternative (and more revealing, though less common) way to look at identification

In our discussion of simultaneous equation models, we talked at length about instrumental variables. In discussing triangular identification, we never mentioned them. It is time to understand what is the link...

Let us consider the first equation in the monetary VAR above. As one can see, there are no contemporaneous relations between variables there. It can therefore be estimated by OLS, as it is basically already a reduced form equation.

The second equation has two variables dated at time t, Y_t and π_t. Potentially we may have a simultaneous equation bias. Y_t may be correlated with $u_{\pi,t}$ It turns out that under the two assumption of a triangular A and diagonal Σ, there is no problem. Indeed, $Cov(Y_t, u_{\pi,t}) = 0$. Going to the third equation, there are three variables at time t. It may look as if we needed two instruments; by the same logic as above, $Cov(Y_t, u_{FFR,t}) = 0$ and $Cov(\pi_t, u_{FFR,t}) = 0$. We can estimate the system with OLS with no problems. An alternative is to use as an instrument $\hat{u}_{Y,t}$. Results are the same.

Impulse response function (IRF)

Given our structural VAR, the most used tool are impulse responses. They are used to study what is the dynamic pattern of the variables in response to shocks.

In general, one can compute IRFs from the VMA representation of the reduced form. From

$$x_t = Cx_{t-1} + v_t$$

the VMA is:

$$x_t = v_t + C^1 v_{t-1} + C^2 v_{t-2} + ... + C^k v_{t-k}$$

recalling that $C = A^{-1}B$ and $v_t = A^{-1}u_t$

$$x_t = A^{-1}u_t + A^{-1}BA^{-1}u_{t-1} + (A^{-1}B)^2 A^{-1}u_{t-2} + ... + (A^{-1}B)^k A^{-1}u_{t-k}$$

These are the structural impulse responses to the structural shocks u_t.

Let us go back to the monetary VAR. The impulse responses to a monetary shock is:

$$\begin{bmatrix} Y_t \\ \pi_t \\ FFR_t \end{bmatrix} = \begin{bmatrix} d_1 & 0 & 0 \\ e_1 & e_2 & 0 \\ f_1 & f_2 & f_3 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ u_{FFR,t} \end{bmatrix}$$

$$\begin{bmatrix} Y_{t+1} \\ \pi_{t+1} \\ FFR_{t+1} \end{bmatrix} = \begin{bmatrix} d_1 & 0 & 0 \\ e_1 & e_2 & 0 \\ f_1 & f_2 & f_3 \end{bmatrix} B \begin{bmatrix} Y_t \\ \pi_t \\ FFR_t \end{bmatrix} = \begin{bmatrix} d_1 & 0 & 0 \\ e_1 & e_2 & 0 \\ f_1 & f_2 & f_3 \end{bmatrix} B \begin{bmatrix} d_1 & 0 & 0 \\ e_1 & e_2 & 0 \\ f_1 & f_2 & f_3 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ u_{FFR,t} \end{bmatrix}$$

$$\begin{bmatrix} Y_{t+2} \\ \pi_{t+2} \\ FFR_{t+2} \end{bmatrix} = \begin{bmatrix} d_1 & 0 & 0 \\ e_1 & e_2 & 0 \\ f_1 & f_2 & f_3 \end{bmatrix} B \begin{bmatrix} Y_{t+1} \\ \pi_{t+1} \\ FFR_{t+1} \end{bmatrix} = \begin{bmatrix} d_1 & 0 & 0 \\ e_1 & e_2 & 0 \\ f_1 & f_2 & f_3 \end{bmatrix}^2 B^2 \begin{bmatrix} d_1 & 0 & 0 \\ e_1 & e_2 & 0 \\ f_1 & f_2 & f_3 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ u_{FFR,t} \end{bmatrix}$$

Notice that the impulse responses will be different according to the identification scheme one chooses. Sims proposed to check many different ordering and see whether results were robust to various orderings.

Forecast error variance decomposition (FEVD)

Another statistic which is often employed is the forecast error variance decomposition.

We have seen above that the variance of the one-step ahead forecast error is Ω.

The variance of the two-step ahead forecast error is

$$\Omega + \Psi_1 \Omega \Psi_1'$$

and the variance of the two-step ahead forecast error i

$$\Omega + \Psi_1 \Omega \Psi_1' + \Psi_2 \Omega \Psi_2'$$

and so on...

One may want to know how much of this variance is due to different orthogonalized shocks. Let us use the Choleski example.

For the one step ahead, the variance due to the $j-th$ shock is: $G_j G_j'$, where G_j is the $j-th$ column of G.

The percentage of variance explained by the $j-th$ shock can be found as

$$diag(G_j G_j')/diag(\Omega)$$

For the two steps ahead, the variance due to the $j-th$ shock is:

$$G_j G_j' + \Psi_1 G_j G_j' \Psi_1'$$

The percentage of variance explained by the $j-th$ shock is

$$diag(G_j G_j' + \Psi_1 G_j G_j' \Psi_1')/diag(\Omega + \Psi_1 \Omega \Psi_1')$$

For the three steps ahead, the variance due to the $j-th$ shock is:

$$G_j G_j' + \Psi_1 G_j G_j' \Psi_1' + \Psi_2 G_j G_j' \Psi_2'$$

The percentage of variance explained by the $j-th$ shock is:

$$diag(G_j G_j' + \Psi_1 G_j G_j' \Psi_1' + \Psi_2 G_j G_j' \Psi_2')/diag(\Omega + \Psi_1 \Omega \Psi_1' + \Psi_2 \Omega \Psi_2')$$

The FEVD will change as identification change.

Non- triangular identification schemes

Triangularity is clearly a very specific assumption. Researchers right after Sims' paper have started to select identification schemes that were not triangular. An example is in Gordon and Leeper (1994)

$$A_0 = \begin{bmatrix} 1 & b_{12} & 0 & 0 & 0 & b_{16} & b_{17} \\ b_{21} & 1 & 0 & b_{24} & b_{25} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & b_{43} & 1 & 0 & 0 & 0 \\ 0 & 0 & b_{53} & b_{54} & 1 & 0 & 0 \\ 0 & 0 & b_{63} & b_{64} & b_{65} & 1 & 0 \\ 0 & 0 & b_{73} & b_{74} & b_{75} & b_{76} & 1 \end{bmatrix}$$

$$x_t = \begin{bmatrix} i \text{ interest rate} \\ m \text{ money} \\ ur \text{ unemp rate} \\ y \text{ real GDP} \\ p \text{ CPI} \\ i_{10} \text{ 10yr bond rate} \\ cp \text{ commodity price index} \end{bmatrix}$$

in which the first equation is interpreted as money supply function and the second as a money demand. The rest of the system has a triangular structure. It is clear that contemporaneous values of ur, y, p, i_{10}, cp can be used as instruments in estimating the first and the second equation.

Long run identification

Blanchard and Quah (1989) have introduced a new kind of identification scheme. From the point of view of the old simultaneous equations systems, the idea of BQ is simply to obtain identification by restricting linear combinations of parameters to equal 0.

The interesting thing is what kind of linear combination they set to zero. They have a structural model whose VMA is:

$$\begin{bmatrix} \Delta y_t \\ u_t \end{bmatrix} = \Psi(L)u_t$$

and their idea is that one of the orthogonal shocks in u_t has no long run effect on the level of y, while the other is allowed to have non-zero effect on the level of y. Write explicitely $\Psi(L)u_t = \begin{bmatrix} \Psi_{11}(L) & \Psi_{12}(L) \\ \Psi_{21}(L) & \Psi_{22}(L) \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$. It is clear that the impulse response of a shock u_{1t} on Δy is given by the $\Psi_{11}(L)$ polynomial. It should also be clear that the long run effect of a shock u_{1t} on the levels of y is given by $\Psi_{11}(1)$.

Therefore the identification proposed by BQ is the one such that $\Psi_{11}(1) = 0$, that is, the "model" they have in mind is one in which u_{1t} a demand shock (which has only a temporary effect on y). No restrictions are put on the other three polynomials in $\Psi(L)$.

How is this implemented? Suppose that we have a VAR(1) with structural form:

$$Az_t = Bz_{t-1} + v_t$$

with reduced form:

$$z_t = Cz_{t-1} + u_t = A^{-1}Bz_{t-1} + A^{-1}v_t$$

We want to select A such that $Cov(v_t) = I$ and the matrix of cumulated long run responses has a zero, say at the position $(1, 2)$.

The VMA for the reduced form will be:

$$z_t = D(L)u_t = (I + CL + C^2L^2 + ...)u_t$$

Evaluated at $L = 1$, $D(1) = (I - C)^{-1}$

The one for the structural model will be:

$$\begin{aligned} z_t &= (A^{-1} + A^{-1}BA^{-1}L + (A^{-1}B)^2A^{-1}L^2 + ...)v_t = \\ z_t &= (I + A^{-1}B + (A^{-1}B)^2 + ...)A^{-1}v_t \end{aligned}$$

The polynomial evaluated at $L = 1$ will be: $(I - A^{-1}B)^{-1}A^{-1} = (I - C)^{-1}A^{-1} = H$.

We need to find 4 elements in A, three will be obtained from:

$$A^{-1}A^{-1\prime} = \Omega$$

The fourth from

$$H_{12} = 0$$

As seen above, one can use an equivalent IV approach. The issue is to understand what instruments to use.

In larger system, in which the structural long run matrix is not diagonal, one has to use different approaches (the best explanation of this is in Shapiro and Watson, 1989, Sources of Business Cycle Fluctuations, NBER Macroeconomics Annual).

Identification through rotations

This is the general approach to identification. One can separate the identification in two steps:

1. Keep assumption that: $V(u_t) = I \quad (\frac{n(n+1)}{2}$ restrictions)
2. Rotate the shocks in the other $\frac{n(n-1)}{2}$ dimensions

From the reduced form $y_t = C(L)\epsilon_t$
with Choleski we can orthogonalize the shocks:

$$y_t = C(L)PP^{-1}\epsilon_t = C(L)Pu_t$$

where P is the Choleski factor of Ω ($\Omega = PP^{'}$) and $V(u_t) = V(P^{-1}u_t) = I$
you can now rotate these orthonormal shocks with:

$$y_t = C(L)PRR'u_t$$

with $RR' = I$ (orthogonal matrix)
Call $R'u_t = v_t$ "structural shocks" and note that: $V(\nu_t) = R'IR = I$

It turns out that R is a function of $\frac{n(n-1)}{2}$ parameters

Let's make an example with $n = 2$

$$R = \begin{bmatrix} cos\, a & sin\, a \\ -sin\, a & cos\, a \end{bmatrix}$$

parameterized by one parameter: $\frac{n(n-1)}{2} = 1$
If $n = 3$:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\, a & sin\, a \\ 0 & -sin\, a & cos\, a \end{bmatrix} \begin{bmatrix} \cos b & \sin b & 0 \\ -sin\, b & \cos b & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos\, c & 0 & sin\, c \\ 0 & 1 & 0 \\ -sin\, c & 0 & cos\, c \end{bmatrix}$$

1 column is parameterized by $(n-1)$ parameters.
A common case: long-run restrictions on $I(1)$ variables.
Restrictions on the long-run matrix of the system(Blanchard-Quah)

Reduced form:

$$\Delta y_t = C\Delta y_{t-1} + \epsilon_t \rightarrow \Delta y_t = C(L)\epsilon_t$$

Structural form:

$$\Delta y_t = \Theta(L)\nu_t \quad V(\nu_t) = I$$

but now $\Theta(1)$ has some zeros.
Blanchard-Quah:

$$z_t = \Theta(L)\nu_t$$

$$\begin{bmatrix} \Delta y_t \\ u_t \end{bmatrix} = \begin{bmatrix} \theta_{11}(L) & \theta_{12}(L) \\ \theta_{21}(L) & \theta_{22}(L) \end{bmatrix} \begin{bmatrix} \nu_{1t} \\ \nu_{2t} \end{bmatrix}$$

ν_{1t}: demand shock has no long-run effect on the level of y
ν_{2t}: supply shock is allowed to have long-run effect on y

$$\Delta y = \theta_{11}(L)\nu_{1t} \qquad y_{t+\infty} = \theta_{11}(1)\nu_{1t} \quad \Rightarrow \quad \theta_{11}(1) = 0$$

Take reduced form:

$$\begin{bmatrix} \Delta y \\ u \end{bmatrix} = C(L)\epsilon_t \qquad V(\epsilon_t) = \Sigma$$

1. Choleski:

$$\begin{bmatrix} \Delta y \\ u \end{bmatrix} = G(L)P\tilde{\nu_t} \qquad V(\tilde{\nu_t}) = I$$

2. Pick the last restriction so to have:

$$\begin{bmatrix} \Delta y \\ u \end{bmatrix} = C(L)PR' \underbrace{R\tilde{\nu}_t}_{\nu_t}$$

$\Rightarrow \Theta(1) = C(1)PR'$ with: $\Theta_{11}(1) = 0$
From

$$\begin{bmatrix} c_{11}(1) & c_{12}(1) \\ c_{21}(1) & c_{22}(1) \end{bmatrix} \begin{bmatrix} P_{11} & 0 \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} cos\, a & sin\, a \\ -sin\, a & cos\, a \end{bmatrix}$$

the restriction on the (1,1) element is: $\begin{bmatrix} c_{11}(1)P_{11} + c_{12}(1)P_{21} & c_{12}(1)P_{22} \end{bmatrix} \begin{bmatrix} cos\, a \\ -sin\, a \end{bmatrix} = 0$.

Only 1 parameter to pick to implement Blanchard-Quah (Matlab: numerical)

Cointegration

Two $I(1)$ (with unit roots) time series, y_t and x_t, are said to be cointegrated if there exist a linear combination, for example,

$$y_t - \alpha x_t$$

which is stationary. The vector $[1 \ -\alpha]$ is a cointegrating vector. Cointegrating vectors are not unique.

In general, given a $(n \times 1)$ vector y_t, the elements in y_t are cointegrated if there exist a $(n \times r)$ matrix β, such that

$$\beta' y_t = z_t \sim I(0)$$

The columns of β are linearly independent ($r < n$ and $\text{rank}(\beta) = r$).

Cointegration relations are often seen as long run equilibrium relations. The variables in z_t represent stationary fluctuations around these long run relations.

There are many economic examples. For example, real GDP is $I(1)$, real consumption is $I(1)$, but the consumption-output ratio is more or less stable in time, that is, $\log(c_t) - \log(y_t)$ is stationary. The cointegrating vector in this case is $[1 \ \ -1]$.

The purchasing power parity. The PPP says that the same good should sell at the same effective price in two countries.

If we denote by P_t the home price, by P_t^* the foreign price and by E_t the nominal exchange rate (home currency per foreign currency), the PPP holds that:

$$P_t = E_t P_t^*$$

in logs:

$$\log(P_t) = \log(E_t) + \log(P_t^*)$$

In practice there may be factors that prevent this equation to hold with equality, so a weaker version is

$$z_t = \log(P_t) - \log(E_t) - \log(P_t^*)$$

is stationary (and $\log(P_t), \log(E_t), \log(P_t^*)$ are $I(1)$). This implies that the three variables are cointegrated with cointegrating vector $[1 \ -1 \ -1]$.

The basic RBC model with unit root labor augmenting productivity.

Take for example the vector $[\log(Y_t), \log(C_t), \log(I_t)]$. The model implies that each of the elements of the vector is $I(1)$ and that $\log(C_t) - \log(Y_t)$ and $\log(I_t) - \log(Y_t)$ are both stationary. Therefore, both $[-1 \ 1 \ 0]$ and $[-1 \ 0 \ 1]$ are cointegrating vectors.

Let us work out few properties from an example:

$$\begin{bmatrix} 1 & -\gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}$$

Inverting the matrix of the contemporaneous relations,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 0 & \gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} 1 & \gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}$$

From the "reduced form", we obtain the forecast errors. The forecast error for $y_{1,t}$ is $\varepsilon_{1,t} = u_{1,t} + \gamma u_{2,t}$. The forecast error for $y_{2,t}$ is $\varepsilon_{2,t} = u_{2,t}$.

The autoregressive matrix has two eigenvalues, one of which is equal to 1 (or, equivalently, one of the roots of $|I - Az| = 0$ is equal to one, try...). There is a unit root: $y_{1,t}$ and $y_{2,t}$ are both $I(1)$.

Note that $y_{1,t} - \gamma y_{2,t} = u_{1,t}$ is stationary. The vector $[1 \ -\gamma]$ is a cointegrating vector. Cointegrating vectors are not unique, $[2 \ -2\gamma]$ is a cointegrating. vector, $[-1 \ \gamma]$ is a cointegrating vector.

Let us derive the Wold representation for Δy_t (as y_t is $I(1)$ a Wold representation does not exists, while it exists forΔy_t).

We need the MA polynomial for Δy_t. As there is a unit root, we cannot obtain it by inverting the VAR(1).

Let us start by writing the two equations in first differences:

$$\begin{aligned} \Delta y_{1,t} &= \gamma \Delta y_{2,t} + \Delta u_{1,t} = \gamma u_{2,t} + \Delta u_{1,t} \\ \Delta y_{2,t} &= u_{2,t} \end{aligned}$$

In matrix form

$$\begin{aligned} \begin{bmatrix} 1 & -\gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{bmatrix} &= \begin{bmatrix} 1-L & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \\ \begin{bmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{bmatrix} &= \begin{bmatrix} 1-L & \gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \end{aligned}$$

The Wold representation has therefore to satisfy:

$$\begin{aligned} \Delta y_{1,t} &= (1-L)u_{1,t} + \gamma u_{2,t} \\ \Delta y_{2,t} &= u_{2,t} \end{aligned}$$

but we need to express Δy_t as a function of forecast errrors (that's the Wold, right?):

$$\begin{aligned} \Delta y_{1,t} &= (1-L)u_{1,t} + \gamma u_{2,t} = (1-L)(\varepsilon_{1,t} - \gamma \varepsilon_{2,t}) + \gamma \varepsilon_{2,t} = (1-L)\varepsilon_{1,t} + \gamma L \varepsilon_{2,t} \\ \Delta y_{2,t} &= \varepsilon_{2,t} \end{aligned}$$

We can rewrite

$$\begin{bmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{bmatrix} = \Psi(L)\varepsilon_t = \begin{bmatrix} 1-L & \gamma L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

If we want to find a VAR for Δy_t we have to invert $\Psi(L)$. It is easy now to check that $|\Psi(z)| = 0$ when $z = 1$. The polynomial $\Psi(L)$ cannot be inverted and $\Psi(1)$ has reduced rank. Therefore a VAR for Δy_t with a finite number of lags does not exist!

Impulse responses. From the VMA for the structural model:

$$\begin{aligned} \begin{bmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{bmatrix} &= \begin{bmatrix} 1-L & \gamma L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \\ &= \begin{bmatrix} 1-L & \gamma L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \\ &= \begin{bmatrix} 1-L & \gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \end{aligned}$$

we can compute the long run response of this system to either a $u_{1,t}$ or a $u_{2,t}$ shock. The response will be 0 for the first differences, but it will be

$$\Psi(1) = \begin{bmatrix} 0 & \gamma \\ 0 & 1 \end{bmatrix}$$

for the levels. A unitary u_1 shock won't have any effect on the long run levels of y_t. A unitary u_2 shock will have an effect on the long run levels: y_1 will be equal to γ, y_2 will be equal to 1. The two variables have both to move in the long run, so to guarantee that a linear combination will be stationary (remember: the long run impact of a shock on a stationary variable is zero). What linear combination? The only one is: $y_{1,t} - \gamma y_{2,t}$. The long run effect of a shock will be: $\gamma - \gamma \cdot 1 = 0$.

Implications for the VAR representation

Note that $\Phi(L) = I - \Phi_1 L = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & \gamma \\ 0 & 1 \end{bmatrix} L = \begin{bmatrix} 1 & -\gamma L \\ 0 & 1-L \end{bmatrix}$

$\Phi(1) = \begin{bmatrix} 1 & -\gamma \\ 0 & 0 \end{bmatrix}$ which has rank 1 (in general, the rank of $\Phi(1)$ in a cointegrated system is k). The matrix $\Phi(1)$ can be written as: BA', where both A and B are $(n \times k)$ matrices. Let the rows of A be the cointegrating vectors, then in our case, $\begin{bmatrix} 1 & -\gamma \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \begin{bmatrix} 1 & -\gamma \end{bmatrix}$ so that: $B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

Stock and Watson common trends representation

Let us write the Beverigde-Nelson representation, from

$$\Psi(L) = \Psi(1) + \Psi^*(L)(1-L)$$

by equating elements, one obtains:

$$\Psi^*(L) = \begin{bmatrix} 1 & -\gamma \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} L + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} L^2 + ...$$

From which:

$$\Delta y_t = \begin{bmatrix} 0 & \gamma \\ 0 & 1 \end{bmatrix} \varepsilon_t + \begin{bmatrix} 1 & -\gamma \\ 0 & 0 \end{bmatrix} (1-L)\varepsilon_t$$

Cumulating:

$$\begin{aligned} y_t &= \begin{bmatrix} 0 & \gamma \\ 0 & 1 \end{bmatrix} \sum_{i=0}^{t} \begin{bmatrix} \varepsilon_{1,i} \\ \varepsilon_{2,i} \end{bmatrix} + \begin{bmatrix} 1 & -\gamma \\ 0 & 0 \end{bmatrix} \varepsilon_t \qquad (6) \\ & \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \sum_{i=0}^{t} \varepsilon_{2,i} + \begin{bmatrix} 1 & -\gamma \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \end{aligned}$$

For the structural model:

$$\begin{aligned} y_t &= \begin{bmatrix} 0 & \gamma \\ 0 & 1 \end{bmatrix} \sum_{i=0}^{t} \begin{bmatrix} \varepsilon_{1,i} \\ \varepsilon_{2,i} \end{bmatrix} + \begin{bmatrix} 1 & -\gamma \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \\ &= \begin{bmatrix} 0 & \gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \gamma \\ 0 & 1 \end{bmatrix} \sum_{i=0}^{t} \begin{bmatrix} u_{1,i} \\ u_{2,i} \end{bmatrix} + \begin{bmatrix} 1 & -\gamma \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & \gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \qquad (7) \\ &= \begin{bmatrix} 0 & \gamma \\ 0 & 1 \end{bmatrix} \sum_{i=0}^{t} \begin{bmatrix} u_{1,i} \\ u_{2,i} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \qquad (8) \\ &= \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \sum_{i=0}^{t} u_{2,i} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_{1,t} \qquad (9) \end{aligned}$$

There is only one random walk that drives the permanent component of the system.

This is true in general: in a cointegrated system with k cointegrating relations, there are $n-k$ common trends ($n-k$ random walks drive the non-stationary part of the system).

Implications from the VMA representation

From the BN decomposition in equation (6), we can also see that:

$$\begin{bmatrix} 1 & -\gamma \end{bmatrix} y_t = \begin{bmatrix} 1 & -\gamma \end{bmatrix} \begin{bmatrix} 0 & \gamma \\ 0 & 1 \end{bmatrix} \sum_{i=0}^{t} \begin{bmatrix} \varepsilon_{1,i} \\ \varepsilon_{2,i} \end{bmatrix} + \begin{bmatrix} 1 & -\gamma \end{bmatrix} \begin{bmatrix} 1 & -\gamma \\ 0 & 0 \end{bmatrix} \varepsilon_t$$

The left hand side is stationary, the right hand side must also be stationary. This implies that: $\underset{(1\times 2)}{a'} \underset{(2\times 2)}{\Psi(1)} = 0$. If there were k cointegrating vectors: $\underset{(k\times n)}{A'} \underset{(n\times n)}{\Psi(1)} = \underset{(k\times n)}{0}$

Vector Error Correction Mechanism (VECM)

Let us take first differences of the process:

$$\begin{aligned}\begin{bmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{bmatrix} &= \begin{bmatrix} -1 & \gamma \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} 1 & \gamma \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \\ &= \begin{bmatrix} -1 & \gamma \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t,1} \\ \varepsilon_{2,t} \end{bmatrix}\end{aligned}$$

This is the representation for Δy_t. The lagged level still appears. As we saw before, this is not a VAR in first differences.

This can also be written as:

$$\begin{bmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix} z_{t-1} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

where: $z_t = y_{1,t} - \gamma y_{2,t}$. Notice if there are no shocks, $\Delta y_{1,t}$ and $\Delta y_{2,t}$ will be zero after some time. The long run equilibrium level of $y_{1,t}$ and $y_{2,t}$ will be $z_t = y_{1,t} - \gamma y_{2,t} = 0$, that is $y_{1,t} = \gamma y_{2,t}$. In the short run, shocks will move $\Delta y_{1,t}$ and $\Delta y_{2,t}$ away from zero and therefore will move z_t as well. If $z_{t-1} = y_{1,t-1} - \gamma y_{2,t-1} > 0$ (there is a positive deviation from the long run equilibrium), what happens in period t? $\Delta y_{1,t} < 0$. That is, the system will move towards the long run equilibrium relation (only $y_{1,t}$ will adjust in this case, as $y_{2,t}$ is a random walk). This is the reason for the name "error-correction". With the idea of cointegration we can model simultaneously the long run dynamics, driven by the common trends, and the short run dynamics with the idea of error correction mechanisms.

Granger representation theorem

We can put together all what we have seen so far.

Consider a $(n \times 1)$ vector y_t, where $\Delta y_t = \Psi(L)\varepsilon_t$. Suppose there are k cointegrating relations.

Then, there exists a $(n \times k)$ matrix A, with linearly independent columns, such that $A'y_t = z_t$ is stationary (the columns of A are cointegrating vectors and the vectors are linearly independent).

The matrix A is such that $A'\Psi(1) = 0$

If the process can be represented as a $VAR(p)$

$$y_t - \Phi_1 y_{t-1} - ... - \Phi_p y_{t-p} = \varepsilon_t$$

then there exists a $(n \times k)$ matrix B such that $\Phi(1) = BA'$ (the rank of $\Phi(1)$ is therefore equal to k. Equivalently, $\det(A(z)) = 0$ has $n - k$ unit roots)

Moreover, there exist matrices $\zeta_1, \zeta_2, ..., \zeta_{p-1}$ such that:

$$\begin{aligned}\Delta y_t &= \zeta_1 \Delta y_{t-1} + \zeta_{p-1} \Delta y_{t-p+1} - BA' y_{t-1} + \varepsilon_t \\ &\quad \zeta_1 \Delta y_{t-1} + \zeta_{p-1} \Delta y_{t-p+1} - B z_{t-1} + \varepsilon_t\end{aligned}$$

(the matrices $\zeta_1, \zeta_2, ..., \zeta_{p-1}$ can be obtained from the Beveridge-Nelson decomposition of the $\Phi(L)$ polynomial)

Remarks

The cointegration rank (the number of linearly independent cointegrating vectors) cannot be larger than $n-1 : k \leq n-1$.

From the VECM:

$$\zeta(L)\Delta y_t = -\Phi(1)y_{t-1} + \varepsilon_t$$

write:

$$\Phi(1)y_{t-1} = -\zeta(L)\Delta y_t + \varepsilon_t$$

The right hand side is stationary, as it is function of Δy_t and ε_t (call the right hand side w_t).

The left hand side must be stationary as well. If the rank of $\Phi(1)$ was n, we could invert it to obtain:

$$y_{t-1} = \Phi(1)^{-1} w_t$$

and y_{t-1} will be stationary, as a linear combination of stationary variables is stationary.

Either one of the two:

- if y_t is non-stationary, $\Phi(1)$ cannot be invertible and therefore it must be $k < n$
- if $\Phi(1)$ is invertible (rank n), y_t has to be stationary

The ECM is not unique (same logic of the fact that cointegrating vectors are not unique).

If we take the above ECM representation

$$\Delta y_t = \zeta_1 \Delta y_{t-1} + \zeta_{p-1}\Delta y_{t-p+1} - BA' y_{t-1} + \varepsilon_t$$

and write:

$$\Delta y_t = \zeta_1 \Delta y_{t-1} + \zeta_{p-1}\Delta y_{t-p+1} - BQ^{-1}QA' y_{t-1} + \varepsilon_t$$

where Q is any invertible $(k \times k)$ matrix, we can write

$$\Delta y_t = \zeta_1 \Delta y_{t-1} + \zeta_{p-1}\Delta y_{t-p+1} - B^* A^{*\prime} y_{t-1} + \varepsilon_t$$

If the rows of A' are cointegration vectors, also the rows of $A^{*\prime}$ will be cointegration vectors.

There is an *identification problem.*

Estimation

If cointegration implies that $y_t - \alpha x_t \sim I(0)$ an intuitive approach is to estimate the regression

$$y_t = \alpha x_t + \varepsilon_t \tag{10}$$

and test if the residuals $\hat{\varepsilon}_t$ are stationary. This is the approach proposed by Engle and Granger (1987).

Stock (1987) showed that if the variables are cointegrated, the OLS estimator of α will converge to the true value at rate T (superconsistency). The only linear combination that makes ε_t stationary (therefore with *constant and finite* variance) is $[1 \quad -\alpha]$. OLS finds a minimum variance residual and therefore will pick up precisely that α that obtains "low variance" (stationary) residuals (of course, in small samples we may have problems).

The test regression is nothing but the DF regression:

$$\Delta\hat{\varepsilon}_t = \phi\hat{\varepsilon}_{t-1} + v_t$$

where the null hypothesis of stationarity of $\hat{\varepsilon}_t$ is $H_0 : \phi = 0$.

Problem: $\hat{\varepsilon}_t$ are generated from the regression above. Since the $\hat{\varepsilon}_t$ have been estimated by minimizing their variance, the test using the DF tables will be bised towards finding stationarity of $\hat{\varepsilon}_t$. The distribution of $\hat{\phi}$ under $H_0 : \phi = 0$ has been tabulated and depends also on the number of regressors in equation (10).

If the residuals $\hat{\varepsilon}_t$ are not white noise, the right approach is to use an ADF test, based on the regression:

$$\Delta\hat{\varepsilon}_t = \phi\hat{\varepsilon}_{t-1} + a(L)\Delta\hat{\varepsilon}_{t-1} + v_t$$

There are tables for this test as well, as $\hat{\varepsilon}_t$ is not observed.

The estimation of the ECM

$$\begin{aligned}
\Delta z_t &= BA' z_{t-1} + \zeta_1 \Delta z_{t-1} + \varepsilon_t \\
&= \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \begin{bmatrix} 1 & -\alpha \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \zeta_1 \Delta z_{t-1} + \varepsilon_t
\end{aligned}$$

cannot be performed equation by equation as there are cross-equation restrictions.

The suggestion of Engle and Granger is to follow a two-step approach, based on the fact that:

$$\begin{aligned}
\Delta z_t &= \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \begin{bmatrix} 1 & -\alpha \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \zeta_1 \Delta z_{t-1} + \varepsilon_t \\
&= \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \hat{\varepsilon}_{t-1} + \zeta_1 \Delta z_{t-1} + \varepsilon_t
\end{aligned}$$

First, one estimates α from $y_t = \alpha x_t + \varepsilon_t$ and test for stationarity of $\hat{\varepsilon}_t$ Second, one estimates the two equations above, which are like a VAR with an exogenous stationary variable.

Limits

- only one cointegration relation, as it is based on the regression $y_t = \alpha z_t + \varepsilon_t$ (z_t can be a vector).
- two step (not efficient).

An alternative approach has been proposed by Stock and Watson and Johansen and it is based on estimating the rank of the matrix $\Phi(1) = BA'$ in

$$\Delta z_t = BA' z_{t-1} + \sum_{i=1}^{p} \zeta_i \Delta z_{t-i} + \varepsilon_t \tag{11}$$

As we have seen above, the rank of $\Phi(1)$ is equal to the number of cointegration relations (what is the implication that $\Phi(1)$ has rank zero? and that it has rank n?).

The test is based on testing the rank of BA' (equation (11) is extremely similar to a multivariate ADF equation).

If you go back to your linear algebra courses, the rank of a matrix is equal to the number of non-zero eigenvalues.

The test will therefore be a test on how many eigenvalues of BA' are equal to zero.

We first obtain an estimate of BA', we compute the eigenvalues, we order them from the largest to smallest ($\lambda_1 > \lambda_2 > ... > \lambda_n$)[4].

If there is no cointegration, $rank(BA') = 0$. Then $\lambda_1 = 0$ (and of course, for any other i). This means that $\ln(1 - \lambda_i)$ will be equal to zero for any i. If $rank(BA') = 1$, then $\ln(1 - \lambda_1) < 0$ and $\ln(1 - \lambda_i) = 0$, for $i > 1$ and so on...

In practice, we base our inference on the two statistics:

$$\begin{aligned} \lambda_{trace}(r) &= -T \sum_{i=r+1}^{n} \ln(1 - \hat{\lambda}_i) \\ \lambda_{\max}(r, r+1) &= -T \ln(1 - \hat{\lambda}_{r+1}) \end{aligned}$$

With $\lambda_{trace}(r)$ one tests $H_0 : \lambda_{r+1} = \lambda_{r+2} = ... = \lambda_n = 0$ against $H_1 :$ at least one of the $\lambda_{r+1}, \lambda_{r+2}, ..., \lambda_n$ is different from zero. If the null is true, $\lambda_{trace}(r)$ will be zero. If the null is false, $\lambda_{trace}(r)$ will be positive and large.

With $\lambda_{\max}(r, r+1)$ one tests $H_0 : \lambda_{r+1} = 0$ (there are r coint.relations) against $H_0 : \lambda_{r+1} \neq 0$ (there are $r + 1$ coint. relations). If the null is true,

[4] Actually, we do not compute the eigenvalues of BA' directly, but we compute the eigenvalues of a "normalized" version of BA'. In this way, we "rescale" the eigenvalues so to make sure that they are all smaller than 1. Keep on reading to understand why.

$\lambda_{\max}(r, r+1)$ will be zero. If the null is false, $\lambda_{\max}(r, r+1)$ will be positive and large.

The distribution here again is non-standard and depends on $n - r$ and on the deterministic components appearing in the cointegrating relations.

How to estimate the parameters in the ECM model? Johansen has developed a ML procedure. We have no time to go into the details, but it suffices to know that it works!!!

Sims, Stock and Watson (Econometrica, 1990, SSW)

This is one of the best econometric papers. Here, I just use some of their results.

The issue is what VAR specification one should use if the DGP is a VAR in levels and the variables are non-stationary $I(1)$.

Suppose one runs a VAR in levels.

The results in SSW shows that all the coefficients in the VAR will be consistently estimated either if the variables are cointegrated or if they are not cointegrated.

Suppose one estimates a VAR in first differences?

if the variables are $I(1)$ but not cointegrated, this is the right thing to do (we saw it above, if the rank of BA' is equal to zero, it means that BA' is a null matrix).

If the variables are cointegrated, it is wrong. The model will be misspecified: we saw above that a VAR system with cointegrated variables cannot be represented as a finite-order VAR in first differences. First differencing will throw away information contained in the cointegrating vectors.

Given the above results, the suggestion of SSW is always run a VAR in levels.
If there is cointegration, this is fine. Of course one can do better by estimating a VECM.

If there is no cointegration, this is fine as well. Of course one can do better by estimating a VAR in first differences (n^2 less parameters to estimate, no problem in the long run when computing impulse responses).

If one runs a VAR in first differences this is wrong if there is cointegration.
One could estimate an ECM, but this requires specifiying the rank of the cointegration space. Given that the cointegration tests has low power, one can aviod the issue of the cointegrating rank, by estimating in levels: still all the coefficients will be consistently estimated.