University of York

Department of Economics

PhD Course 2006

VAR ANALYSIS IN MACROECONOMICS

Lecturer: Professor Mike Wickens

Lecture 1

VAR analysis for stationary data

Contents

Univariate time series processes

x_t $\{t = 1, ..., T\}$ is sample of T observations from a time series process with pdf $f(x_1, ...x_T)$

Properties

mean=$E(x_t) = \mu$

$var(x_t) = E(x_t - \mu)^2 = \gamma(0)$

$cov(x_t, x_{t-s}) = E(x_t - \mu)(x_{t-s} - \mu) = \gamma(s)$ $\quad s = \pm 1, \pm 2, ...$

$cor(x_t, x_{t-s}) = \frac{\gamma(s)}{\gamma(0)} = \rho(s)$, $0 \leq \rho(s) \leq 1$

$\gamma(s)$ is called the autocovariance function

$\rho(s)$ is called the autocorrelation function

Weak stationarity (or covariance stationarity)

x_t is called weakly stationary if $E(x_t)$, $E(x_t - \mu)(x_{t-s} - \mu)$ exist and do not depend on t.

Strong stationarity

x_t is strongly stationary if $f(x_1, ...x_T) = f(x_{T+s+1}, ...x_{2T+s})$ i.e. the whole distribution, and not just the first two moments, are independent of time.

Point estimates of μ and $\gamma(s)$

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

$c(s) = \hat{\gamma}(s) = \frac{1}{T-s} \sum_{t=s+1}^{T} (x_t - \bar{x})(x_{t-s} - \bar{x})$

$r(s) = \hat{\rho}(s) = \frac{c(s)}{c(0)}$

The estimates are consistent if the process is ergodic (approximately, strongly stationary)

A plot of $r(s)$ against s is called the *correlogram*

White noise process

If x_t is stationary, $E(x_t) = 0$, $E_t(x_t^2) = \sigma^2$ and $E(x_t x_{t-s}) = 0$ for $t \neq s$ then x_t is called a white noise process.

We will denote this process e_t. It is also called an $i.i.d(0, \sigma^2)$ process, meaning that each e_t is independently and identically distributed.

Lag operator (backward shift operator)

$L^s x_t = x_{t-s}$

Lead operator (forward shift operator)

$L^{-s} x_t = x_{t+s}$

Some univariate time series models

We assume that $E(x_t) = 0$ in the next sections.

If $E(x_t) = \mu \neq 0$ then we replace x_t below by $x_t - \mu$.

1. Moving average process - MA(q)

$$\begin{aligned} x_t &= e_t + \beta_1 e_{t-1} ... + \beta_q e_{t-q} \\ &= [1 + \beta_1 L ... + \beta_q L^q] e_t \\ &= \beta(L) e_t \end{aligned}$$

is called a q^{th} order MA process and written MA(q).

Properties

$$E(x_t) = E(e_t) + \beta_1 E(e_{t-1}) ... + \beta_q E(e_{t-q}) = 0$$

$$\begin{aligned} V(x_t) &= E(x_t - E(x_t))^2 = E(x_t^2) \\ &= E[\textstyle\sum_{s=0}^{q} \beta_s^{2s} e_{t-s}^2 + 2 \sum_{s=0}^{q} \sum_{r=s}^{q} \beta_s \beta_r e_{t-s} e_{t-r}] \\ &= (\textstyle\sum_{s=0}^{q} \beta_s^{2s}) \sigma^2 \end{aligned}$$

2. Autoregressive models - AR(p)

$$\begin{aligned} x_t + \alpha_1 x_{t-1} ... + \alpha_p x_{t-p} &= e_t \\ [1 + \alpha_1 L ... + \alpha_p L^p] x_t &= e_t \\ \alpha(L) x_t &= e_t \end{aligned}$$

is called a p^{th} order autoregressive process and is written AR(p).

Properties

For convenience, consider an AR(1) process

$$x_t = \alpha x_{t-1} + e_t$$

(i) Invertibility

An AR process can be inverted and written as an MA process provided the roots of the auxiliary equation $\alpha(L) = 0$ are all greater than unity in absolute value.

For the AR(1) there is one root. It is

$$\begin{aligned} \alpha(L) &= 1 - \alpha L = 0 \\ L &= \frac{1}{\alpha} \end{aligned}$$

Invertibility therefore requires that $|\frac{1}{\alpha}| > 1$ or $|\alpha| < 1$.

In this case, by successive substitution we can write x_t as the MA(∞)

$$\begin{aligned} x_t &= e_t + \alpha e_{t-1} + \alpha^2 e_{t-2} + ... \\ &= \textstyle\sum_{s=0}^{\infty} \alpha^s e_{t-s} \end{aligned}$$

or

$$\begin{aligned} \alpha(L)x_t &= (1-\alpha L)x_t = e_t \\ x_t &= \frac{e_t}{1-\alpha L} \\ &= (\textstyle\sum_{s=0}^{\infty} \alpha^s L^s)e_t \\ &= \textstyle\sum_{s=0}^{\infty} \alpha^s e_{t-s} \end{aligned}$$

Note that $|\alpha| < 1$ guarantees that $\lim_{s\to\infty} \alpha^s = 0$, the key requirement for invertibility.

These results can be generalised to higher order processes. For example, an AR(p) with p distinct roots can be factorised as

$$\alpha(L) = (1-\alpha_1 L)(1-\alpha_2 L)...(1-\alpha_p L)$$

For invertibility - into an MA(∞) - we require that $|\alpha_i| < 1$ for all i, i.e. all of the roots are greater than unity in absolute value. Similarly, an MA(q) process can be inverted to an AR(∞) provided the roots of $\beta(L) = 0$ are all greater than unity in absolute value.

(ii) Variance, autocovariance and autocorrelation

Assuming invertibility we can show that

$$\begin{aligned} E(x_t) &= 0 \\ V(x_t) &= (\textstyle\sum_{s=0}^{\infty} \alpha^{2s})\sigma^2 = \frac{\sigma^2}{1-\alpha^2} \end{aligned}$$

$$\begin{aligned} cov(x_t, x_{t-1}) &= E(x_t, x_{t-1}) \\ &= \alpha E(x_{t-1}^2) + E(x_{t-1}e_t) \\ &= \alpha\frac{\sigma^2}{1-\alpha^2} \end{aligned}$$

$$\begin{aligned} cov(x_t, x_{t-s}) &= E(x_t, x_{t-s}) \\ &= \alpha^s \frac{\sigma^2}{1-\alpha^2} \end{aligned}$$

Hence

$$\rho(s) = \alpha^s$$

This tells us how the autocorrelation function would look for an AR(1).

We note that the autocorrelation function coefficients are also the MA coefficients obtained by inverting the AR process. This is true more generally. The coefficients of the AR process are called the *partial* autocorrelation function. The partial autocorrelation function of an MA process is obtained by inverting it to an AR process.

3. ARMA(p,q) models

A more general model combines the AR(p) and the MA(q) to give

$$\alpha(L)x_t = \beta(L)e_t$$

Again this can be written as either an AR(∞) or an MA(∞) provided the invertibility conditions are satisfied.

In practice the advantage of the ARMA is that much smaller values of p and q will represent x_t than for either an AR or an MA.

4. Non-stationary models

Consider the AR(1) with $\alpha = 1$, i.e. a unit root process,

$$\begin{aligned} x_t &= x_{t-1} + e_t \\ \Delta x_t &= e_t \end{aligned}$$

This known as a random walk

If e_t is not an $i.i.d(0, \sigma^2)$ process, for example if $E(e_t^2)$ is not constant, but has zero mean, then x_t is known as a Martingale process.

The properties of x_t are more complicated

$$\begin{aligned} x_t &= x_0 + e_t + e_{t-1} + e_{t-2} + ...e_1 \text{ for } x_0 \text{ a fixed number} \\ &= \textstyle\sum_{s=0}^{\infty} e_{t-s} \text{ otherwise} \\ E(x_t) &= 0 \text{ in both cases provided } x_0 = 0 \end{aligned}$$

Assuming that $x_0 = 0$,

$$E(x_t^2) = E(\textstyle\sum_{s=0}^{t-1} e_{t-s})^2 = \sum_{s=0}^{t-1} E(e_{t-s}^2) = t\sigma^2$$

As this depends on t, it violates the assumption of covariance stationarity.

Note: $\lim_{t\to\infty} E(x_t^2) = \infty$. Thus the variance of random walk is unbounded.

In effect x_t needs to be first differenced to become stationary.

x_t is therefore also known as an I(1) process.

Another implication of x_t being I(1) is that any shock e_t has permanent effect on x_t, i.e. it never disappears, or dies away over time. When x_t is I(0) a shock is temporary and does die away over time, i.e. it has a transitory effect.

6. Random walk with drift μ

$$\begin{aligned} \Delta x_t &= \mu + e_t \\ x_t &= x_0 + \mu t + e_t + e_{t-1} + e_{t-2} + ...e_1 \text{ for } x_0 \text{ a fixed number} \end{aligned}$$

Thus, x_t has a linear trend.

In effect x_t consists of the sum of two processes:

a random walk without drift and a linear trend.

$$\begin{aligned} x_t &= z_t + x_0 + \mu t \\ \Delta z_t &= e_t, \quad z_0 = 0 \end{aligned}$$

The linear trend will "dominate" the random walk.

7. ARIMA(p,d,q) processes

$$\alpha(L)\Delta^d x_t = \beta(L)e_t$$

Δ^d denotes that x_t needs to be differenced d times to make x_t stationary. Thus x_t is an I(d) process.

Usually, $d = 1$ is sufficient. In practice, $d = 2$ is likely to be the maximum required in economics - possibly for the price level and the money supply, or more generally for nominal variables.

8. Fractional Integration

There is no necessity for d to be an integer, it can be a fraction, typically with $0 < d < 1$. If $d \geq \frac{1}{2}$ then the process is non-stationary.

The aim of fractional integration is to capture long memory parsimoniously, i.e. economising on the number of parameters. Fractionally integrated processes are popular among time series analysts, but not among economists as there is no obvious economic justification for them.

In general

$$\begin{aligned} (1-L)^d x_t &= (1 - dL + \frac{d^2}{2!}L^2 - \frac{d^3}{3!}L^3 + ...)x_t \\ &= \alpha(L)x_t \end{aligned}$$

Hence, using just one parameter it is possible to represent an AR model with a long distributed lag. The issue is whether the data satisfy such a restriction.

Multivariate time series processes

Let $x_t = (x_{1t}, x_{2t}, \quad , x_{nt})'$ and $e_t = (e_{1t}, e_{2t}, \quad , e_{nt})'$ be n-dimensional vectors with

$$\begin{aligned} E(e_t) &= 0 \\ E(e_t e_t') &= \Sigma \\ E(e_t e_{t-s}') &= 0 \quad s \neq 0 \end{aligned}$$

1. Vector moving average process - VMA(q)

$$\begin{aligned} x_t &= e_t + B_1 e_{t-1} ... + B_q e_{t-q} \\ &= [1 + B_1 L ... + B_q L^q] \mathrm{e}_t \\ &= B(L) e_t, \quad B_0 = I_n \end{aligned}$$

where B_s are $n \times n$ matrices.

$$\begin{aligned} E(x_t) &= 0 \\ V(x_t) &= B(1) \Sigma B(1)' \end{aligned}$$

2. Vector autoregressive process (VAR)

$$\begin{aligned} x_t + A_1 x_{t-1} ... + A_p x_{t-p} &= e_t \\ [1 + A_1 L ... + A_p L^p] x_t &= e_t \\ A(L) x_t &= e_t, \quad A_0 = I_n \end{aligned}$$

We can write the VAR(1) as

$$x_t = A x_{t-1} + e_t$$

3. VARMA process

$$\begin{aligned} x_t + A_1 x_{t-1} ... + A_p x_{t-p} &= e_t + B_1 e_{t-1} ... + B_q e_{t-q} \\ A(L)x_t &= B(L)e_t \end{aligned}$$

Some properties of the VAR

1. In general x_{it} and $x_{j,t-s}$ will be correlated for all $\{i, j\}$ and $\{t, s\}$

2. In general e_{it} and e_{jt} will also be correlated.

3. In general a shock to e_{jt} will affect $x_{i,t+s}$.

4. Impulse response function (IRF)

This is defined as the reponse of $x_{i,t+s}$ to e_{jt}.

The IRF can be obtained in three ways.

(i) Recursive substitution in the VAR

x_{t+1}, x_{t+2},... can be written

$$\begin{aligned} x_{t+1} &= -A_1 x_t - ... - A_p x_{t-p+1} + e_{t+1} \\ &= (A_1^2 - A_2)x_{t-1} + (A_1 A_2 - A_3)x_{t-2} + ... + A_1 A_p x_{t-p} - A_1 e_t + e_{t+1} \\ x_{t+2} &= -A_1 x_{t+1} - ... - A_p x_{t-p+2} + e_{t+2} \\ &= -A_1[-A_1 x_t - ... - A_p x_{t-p+1} + e_{t+1}] - A_2 x_t - ... - A_p x_{t-p+2} + e_{t+2} \\ &= (A_1^2 - A_2)x_t + (A_1 A_2 - A_3)x_{t-1} + ... + A_1 A_p x_{t-p+1} - A_1 e_{t+1} + e_{t+2} \\ &= [-(A_1^2 - A_2)A_1 + (A_1 A_2 - A_3)]x_{t-1} + ... + (A_1^2 - A_2)e_t - A_1 e_{t+1} + e_{t+2} \\ &\quad etc \end{aligned}$$

Hence,

$$\begin{aligned}
\frac{\partial x_{t+1}}{\partial e_t} &= -A_1 \\
\frac{\partial x_{t+2}}{\partial e_t} &= -A_1 \frac{\partial x_{t+1}}{\partial e_t} - A_2 \\
&= A_1^2 - A_2 \\
&\quad etc
\end{aligned}$$

(ii) Inverting the VAR to a VMA

Pre-multiplying the VAR by $A(L)^{-1}$ - assuming that the inverse exists - we obtain a VMA in $B(L)$ that is of infinite length, i.e. a VMA(∞)

$$\begin{aligned}
x_t &= A(L)^{-1} e_t \\
&= B(L) e_t
\end{aligned}$$

The impulse response function can now be written as

$$\frac{\partial x_{i,t+s}}{\partial e_{jt}} = \{B_s\}_{ij}$$

(iii) Using the companion form

In the special case of a VAR(1), the VMA is easily obtained by successive substitution or by inversion.

$$\begin{aligned}
x_t &= A x_{t-1} + e_t \\
&= \textstyle\sum_{s=0}^{\infty} A^s e_{t-s} \\
\frac{\partial x_{t+s}}{\partial e_t} &= A^s
\end{aligned}$$

The companion form converts a VAR(p) into a VAR(1) and hence makes it is possible to exploit this special case.

Companion form

If x_t is an $n \times 1$ vector defined by the VAR(p)

$$A(L)x_t = e_t$$

then, if we define the $np \times 1$ vector $z_t' = (x_t', ..., x_{t-p+1}')$, the VAR(p) can also be written as

$$z_t = \mathcal{A}z_{t-1} + u_t$$

where

$$\mathcal{A} = \begin{bmatrix} -A_1 & -A_2 & .. & -A_{p-2} & -A_{p-1} \\ I_n & 0 & .. & 0 & 0 \\ 0 & I_n & .. & 0 & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & .. & I_n & 0 \end{bmatrix} \quad u_t = \begin{bmatrix} e_t \\ 0 \\ 0 \\ .. \\ 0 \end{bmatrix}$$

It follows that

$$\begin{aligned} z_t &= \mathcal{A}z_{t-1} + u_t \\ &= \textstyle\sum_{s=0}^{\infty} \mathcal{A}^s u_{t-s} \\ \frac{\partial z_{t+s}}{\partial u_t} &= \mathcal{A}^s \end{aligned}$$

Since $x_t = (I, 0, ..., 0)z_t$ and $e_t = (I, 0, ..., 0)u_t$,

$$\begin{aligned} \frac{\partial x_{t+s}}{\partial e_t} &= \frac{(I, 0, ..., 0)\partial z_{t+s}}{(I, 0, ..., 0)\partial u_t} \\ &= (I, 0, ..., 0)\mathcal{A}^s(I, 0, ..., 0)' \end{aligned}$$

Individual variables and shocks can be selected by defining the vector $\ell_i = (0,...,1,...,0)$ which is zero except fpr the i^{th} element. Thus

$$\frac{\partial x_{i,t+s}}{\partial e_{j,t}} = \ell_i \mathcal{A}^s \ell_j'$$

5. Forecasting with a VAR(1)

s-step ahead forecast

The aim is to forecast x_{t+s} using information available at time t. Consider a VAR(1). The solutions for $x_{t+1}, x_{t+2}, ..., x_{t+s}$ are

$$\begin{aligned} x_{t+1} &= Ax_t + e_{t+1} \\ x_{t+2} &= A^2x_t + e_{t+1} + Ae_t \\ x_{t+s} &= A^sx_t + \textstyle\sum_{i=0}^{s-1} A^i e_{t+s-i} \end{aligned}$$

Hence the best forecast conditional on information at time t is

$$E_t x_{t+s} = A^s x_t$$

When A is unknown, it is replaced with an estimate.

For a VAR(p), we construct the companion form and then extract the forecast for x_{t+s} from this.

s-step ahead forecast error

For A known, this is

$$x_{t+s} - E_t x_{t+s} = \textstyle\sum_{i=0}^{s-1} A^i e_{t+s-i}$$

Hence the variance of the forecast error is

$$\begin{aligned} V(x_{t+s}) &= E(x_{t+s} - E_t x_{t+s})(x_{t+s} - E_t x_{t+s})' \\ &= \textstyle\sum_{i=0}^{s-1} A^i \Sigma A^{i\prime} \end{aligned}$$

For A unknown, we must take account of the additional error due to replacing A with an estimate. Most computer programmes do not do this; they treat A as though it were known.

For a VAR(p) we again use the companion form.

s-step ahead forecast error variance decomposition

An issue of interest is how much of the forecast error variance can be attributed to each of the shocks in the VAR. In particular, how much is due to its "own" shock and how much to other shocks? To answer this question we invert the VAR as an VMA and express the forecast error in terms of the VMA. Thus

$$\begin{aligned} x_{t+s} &= B(L)e_{t+s} \\ &= \textstyle\sum_{i=0}^{\infty} B_i e_{t+s-i} \end{aligned}$$

The forecast is then

$$E_t x_{t+s} = \textstyle\sum_{i=s}^{\infty} B_i e_{t+s-i}$$

the forecast error is

$$x_{t+s} - E_t x_{t+s} = \textstyle\sum_{i=0}^{s-1} B_i e_{t+s-i}$$

and the variance of the forecast error is

$$V(x_{t+s}) = V(x_{t+s} - E_t x_{t+s}) = \textstyle\sum_{i=0}^{s-1} B_i \Sigma B_i'$$

The forecast error variance for x_{kt} is given by

$$V(x_{k,t+s}) = \textstyle\sum_{i=0}^{s-1} \{B_i \Sigma B_i'\}_{kk}$$

In general, this will depend on contributions from all of the shocks e_t, and not just on its own shock, e_{kt}. Only if the shocks are uncorrelated, and so Σ is diagonal, will the own shock be the only source of forecast error. In this case

$$V(x_{k,t+s}) = \textstyle\sum_{i=0}^{s-1} \{B_i\}_{kk}^2 \sigma_{kk}$$

Nonetheless, we may wish to know how much of the forecast error variance is due to its own shock and how much to the others. And we may wish to know this for all forecast horizons

$s = 1, 2, 3, ...$ The way to do this is to transform the other shocks so that they are orthogonal to the own shock. Thus we re-write the VMA as

$$\begin{aligned} x_t &= B(L)Q^{-1}Qe_t \\ &= C(L)\varepsilon_t \end{aligned}$$

where Q is chosen so that

$$E(\varepsilon_t \varepsilon_t') = diagonal\{\sigma_{11}, ..., \sigma_{nn}\}$$

It follows that

$$V(x_{k,t+s}) = \sum_{i=0}^{s-1} \{C_i\}_{kk}^2 \sigma_{kk}$$

This can be achieved by using a Choleski decomposition - which makes Q lower triangular - with x_{kt} ordered first. (We will study Choleski decomposition in detail in later lectures.) We can then obtain the *additional* contribution of the other factors by subtracting this expression from the original forecast error variance of x_{kt}.

Innovation accounting

A plot of the forecast error variance decomposition at each horizon and the impulse response function comprise what is known as innovation accounting.

A note on re-basing VAR forecasts

Suppose that we have estimated a VAR and used it to forecast $s-$periods ahead. After the first period of the forecast horizon has passed, should we continue to use the original VAR forecasts or should we re-base the forecasts?

More precisely, if we estimate the VAR over $t = 1, ..., T$ and then forecast periods $T+1, ..., T+s$; in period $T+1$ should we use the original forecasts for periods $T+2, ..., T+s$, or should we re-base the forecasts by using x_{T+1} as the starting values?

Consider the AR(1)

$$x_t = \alpha x_{t-1} + e_t$$

The forecast for $t + s$ is

$$E_t x_{t+s} = \alpha^s x_t$$

Hence the mean-square forecast error is

$$\begin{aligned} MSE(s) &= E_t[x_{t+s} - E_t x_{t+s}]^2 \\ &= E[\textstyle\sum_{i=0}^{s-1} \alpha^i e_{t+i}]^2 \\ &= \sigma^2 \textstyle\sum_{i=0}^{s-1} \alpha^{2i} \\ &= \sigma^2 \dfrac{1-\alpha^{2s}}{1-\alpha^2} \end{aligned}$$

Thus

$$\frac{\partial MSE(s)}{\partial s} = -2\sigma^2 \ln \alpha \frac{\alpha^{2s}}{1-\alpha^2} > 0 \text{ for } 0 < \alpha < 1$$

It follows that the MSE increases the longer the horizon.

For example.

$$\begin{aligned} MSE(1) &= \sigma^2 \\ MSE(2) &= \sigma^2(1+\alpha^2) \end{aligned}$$

The implication is that it is always best to reduce the horizon by re-basing the forecast on the most recent starting values.

And, since $\frac{\partial^2 MSE(s)}{\partial s^2} < 0$, the greatest gains are for s small, i.e. over short horizons.

Estimation of the VAR

$$A(L)x_t = e_t$$

This can be re-written as

$$x_t = Bz_t + e_t$$

where $z_t = \begin{bmatrix} x_{t-1} \\ . \\ x_{t-p+1} \end{bmatrix}$ and $B = [-A_1, ..., -A_{p-1}]$

Suppose that e_t is assumed to be $NID(0, \Sigma)$

i.e. e_t is serially independent and each e_t has a $N(0, \Sigma)$ distribution.

then

$$f(x_t|x_{t-1}, ..., x_1) = f(e_t) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}e_t'\Sigma^{-1}e_t}$$

and

$$\begin{aligned} f(x_1, ..., x_T) &= \prod_{t=1}^{T} f(x_t|x_{t-1}, ..., x_1) = \prod_{t=1}^{T} f(e_t) \\ &= \frac{1}{(2\pi)^{\frac{Tn}{2}}|\Sigma|^{\frac{T}{2}}} e^{-\frac{1}{2}\sum_{t=1}^{T} e_t'\Sigma^{-1}e_t} \\ &= \frac{1}{(2\pi)^{\frac{Tn}{2}}|\Sigma|^{\frac{T}{2}}} e^{-\frac{1}{2}\sum_{t=1}^{T} x_t'A(L)'\Sigma^{-1}A(L)x_t} \end{aligned}$$

The log-likelihood is therefore

$$\begin{aligned} \ln L &= -[\frac{Tn}{2}\ln(2\pi) + \frac{T}{2}\ln|\Sigma| + \frac{1}{2}\sum_{t=1}^{T} x_t'A(L)'\Sigma^{-1}A(L)x_t] \\ &= -[\frac{Tn}{2}\ln(2\pi) + \frac{T}{2}\ln|\Sigma| + \frac{1}{2}\sum_{t=1}^{T}(x_t - Bz_t)'\Sigma^{-1}(x_t - Bz_t)] \end{aligned}$$

Maximising with respect to Σ gives

$$\frac{\partial \ln L(B, \Sigma)}{\partial \Sigma^{-1}} = \frac{T}{2}\Sigma - \frac{1}{2}\sum_{t=1}^{T}(x_t - Bz_t)(x_t - Bz_t)'$$

Setting this to zero and solving gives the MLE estimator of Σ as

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(x_t - Bz_t)(x_t - Bz_t)'$$

Substituting this back into the likelihood function gives the concentrated likelihood

$$\ln L(B, \hat{\Sigma}) = -[\frac{Tn}{2}\ln(2\pi) + \frac{T}{2}\ln|\hat{\Sigma}| + \frac{Tn}{2}]$$

Maximising this with respect to B is identical to minimising $\ln|\hat{\Sigma}|$ with respect to B.

$$\begin{aligned}\frac{\partial \ln|\hat{\Sigma}|}{\partial B} &= \frac{\partial \ln|\hat{\Sigma}|}{\partial \hat{\Sigma}}\frac{\partial \hat{\Sigma}}{\partial B} \\ &= 2\hat{\Sigma}^{-1}\sum_{t=1}^{T}(x_t - Bz_t)z_t'\end{aligned}$$

Setting this to zero and solving for B gives the MLE of B as

$$\hat{B}' = (\sum_{t=1}^{T} z_t z_t')^{-1}(\sum_{t=1}^{T} x_t z_t')$$

Note

1. This is just the multivariate least squares estimator of $x_t = Bz_t + e_t$
2. It does not involve Σ.
3. It can be calculated by applying OLS to each equation separately.

To see this re-write each equation of the VAR as

$$x_i = Z\beta_i + e_i$$

where

$$x_i = \begin{bmatrix} x_{i1} \\ . \\ x_{iT} \end{bmatrix}, \quad Z = \begin{bmatrix} z_1 \\ . \\ z_T \end{bmatrix}, \quad e_i = \begin{bmatrix} e_{i1} \\ . \\ e_{iT} \end{bmatrix}$$

and β_i' is the i^{th} row of B.

$$\mathbf{x} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ . \\ x_T \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} Z & ... & 0 \\ .. & .. & .. \\ 0 & ... & Z \end{bmatrix} = I_n \otimes Z, ...\mathbf{e} = \begin{bmatrix} e_1 \\ . \\ e_T \end{bmatrix}, ...\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ . \\ \beta_n \end{bmatrix}$$

$\otimes$ denotes a Kronecker product.

Hence $\mathbf{e}$ is $N(0, \Omega)$ where $\Omega = \Sigma \otimes I_T$.

Generalised least estimation again gives the MLE.

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{Z}'\Omega^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\Omega^{-1}\mathbf{x} \\
&= (\Sigma^{-1} \otimes Z'Z)^{-1}(\Sigma^{-1} \otimes Z')\mathbf{x} \\
&= (I_n \otimes (Z'Z)^{-1}Z'))\mathbf{x} \\
&= \begin{bmatrix} (Z'Z)^{-1}Z'x_1 \\ \\ . \\ \\ (Z'Z)^{-1}Z'x_n \end{bmatrix}
\end{aligned}$$

Thus

$$\begin{aligned}
\hat{\beta}_i &= (Z'Z)^{-1}Z'x_i \\
V(\hat{\beta}_i) &= \sigma_i^2(Z'Z)^{-1}
\end{aligned}$$

i.e. OLS applied to each equation separately.

This is only true if Σ is unrestricted.

Selecting the lag length of the VAR

This is a major practical problem in VAR analysis.

In general, in econometric modelling there is a trade-off between

- the biases due to omitting variables that should be in

- the inefficiencies (i.e. inaccurate coefficient estimates) due to costs building estimating coefficients that are really zero

To avoid the first we start with a general model and carry out specification tests.

- this give rise to the problem of pre-test bias as the size of the test (the probability of rejecting the correct model) increases after each test. Need to correct for this.

To avoid the second we start with a simple model and carry out misspecification test. In general the problem with this is that it is not clear how the model should be re-specified when the model is rejected.

Using the first approach to choose the lag length of a VAR, start with a general model and then either

(i) find the lag length that maximises R^2.

The problem with this is that $E(R^2) = k$, the number of explanatory variables - here the order of the lag. As $E(\bar{R}^2) = 0$, this is why we should use $\bar{R}^2$ instead.

(ii) find the maximum lag length by testing each equation separately using standard t-tests.

Choose for the whole model the maximum lag in any equation.

There are many problems with this, not least that of correcting for the size of the test due to pre-test bias, i.e. previous decisions (data-snooping) affects the size of the type I error.

The outcome is usually a model with too high a lag and with a large number of insignificant coefficients at lower lags.

The aim of a VAR is to approximate the x_t process.

It has been shown that for most purposes it is better to have a model with as small a number of lags as possible. This can be accomplished by having a very strict significance criterion (i.e. very small size of test).

There is even a danger in this. Consider a dynamic that cyclical behaviour.As the cycle nears zero the coefficients will become insignificantly different from zero. Is it really correct, therefore, to set them to zero or, even worse, to truncate the dynamic and assume that from this point onwards all of the coefficients are zero?

Information criteria

An alternative way of selecting the model is to use an information criterion such as AIC, Schwarz or HQ. If k = number or parameters estimated. Like $\bar{R}^2$, they all aim to penalise the number of coefficients in the model. The aim is to choose k to minimise the criterion.

Akaike information criterion (AIC)

$$\begin{aligned} AIC(k) &= \min_k \{\ln \sigma_i^2 + 2\frac{k}{T}\} \text{ for a single equation} \\ &= \min_k \{\ln |\Sigma| + 2\frac{k}{T}\}, \quad \text{for } n \text{ equations} \end{aligned}$$

Schwarz's Bayesian information criterion (or just SIC, but sometimes called BIC)

$$\begin{aligned} SIC(k) &= \min_k \{\ln \sigma_i^2 + \frac{k \ln T}{T}\} \text{ for a single equation} \\ &= \min_k \{\ln |\Sigma| + \frac{k \ln T}{T}\}, \quad \text{for } n \text{ equations} \end{aligned}$$

Hannan and Quin (HQ) - for $\theta > 2$

$$\begin{aligned} HQ(p) &= \min_k \{\ln \sigma_i^2 + \theta k \frac{\ln(\ln T)}{T}\} \text{ for a single equation} \\ &= \min_k \{\ln |\Sigma| + \theta k \frac{\ln(\ln T)}{T}\}, \quad \text{for } n \text{ equations} \end{aligned}$$

One criterion can be used throughout, or each criterion can be computed and compared the over all results compared. As they differ only in the term penalising the number of coefficients estimated it is possible to guess how the criteria will compare among themselves.

A comparison of the three criteria turns on the degrees of freedom correction. Thus, for any value of k, the AIC criterion will tend to be lower than the SIC criterion if $T > 8$ (i.e. $\ln T > 2$).

Similarly, HQ>AIC if $T \geqslant \exp(\exp(\frac{2}{\theta}))$.

- for $\theta = 1$, $T > 1618$, $\theta = 2$, $T > 15$ etc

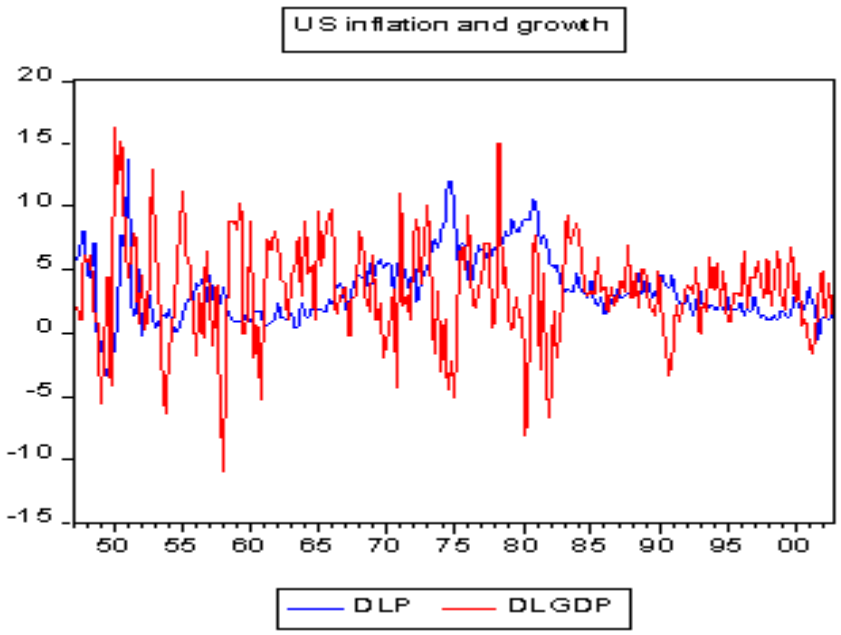
.

Recall the aim is to choose k to minimise the criteria and the criterion that tends to give the lowest value is the AIC. However, the AIC criterion still tends to select a model with too many parameters. So the information criteria, although widely used, are not that helpful.

Numerical Example

To illustrate these ideas we consider a two variable VAR in inflation (Δp_t) and growth (Δy_t), where p_t=log price US level and y_t=log US real GDP. The data are monthly from 1947.1 -2002.4. All estimation is carried out in EViews4.

The data are



The ADF statistics for Δp_t and Δy_t are -3.77 and -10.38. The lag length for Δp_t was one, and for $\Delta y_t$was zero; both were based on the SIC. We can therefore reject the null hypothesis that they are I(1) in favour of I(0).

The correlation between the two series is -0.155.

The autocorrelation and partial autocorrelation functions show that Δy_t has almost no memory, but Δp_t has a very long memory. In fact, although not reported, the autocorrelations are significant even at a 36 month lag.

	AC	PAC	AC	PAC
1	0.795	0.795	0.342	0.342
2	0.715	0.223	0.189	0.081
3	0.648	0.084	-0.005	-0.105
4	0.522	-0.161	-0.115	-0.115
5	0.477	0.080	-0.172	-0.098
6	0.399	-0.047	-0.100	0.017
7	0.377	0.125	-0.084	-0.031
8	0.407	0.178	-0.046	-0.028
9	0.384	0.008	0.052	0.067
10	0.434	0.151	0.063	0.016
11	0.418	-0.073	0.029	-0.036
12	0.420	0.061	-0.126	-0.181

Inspection of the AC and PAC give us an idea of the values of p and q that we should choose if we wish to model a process as an ARMA(p,q). The PAC for Δy_t and the geometric decline of the AC suggest that we need an AR(1). Further estimation and tests show that an AR(1) does fit Δy_t best and has an $\bar{R}^2 = 0.118$. The process generating Δp_t is clearly more complicated. The PAC suggests p should be at least 2. An AR(4) is the best fit and has $\bar{R}^2 = 0.665$.

We now consider a VAR in Δp_t and Δy_t

The lag selection criteria starting witha maximum lag of 12 are

The asterisk denotes the lag length selected by each criterion. The information criteria suggest a lag length of 2, whilst the likelihood ratio test indicates 5. We note, however, that in the inflation equation the maximum significant lag is 10 (for Δp_{t-10}).

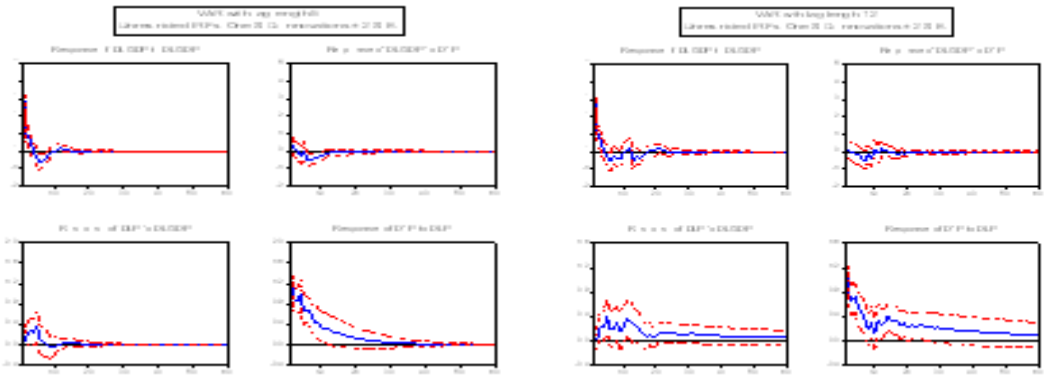
The problem is that including insignificant lags reduces the efficiency of the coefficient estimates (but does not bias them) and increases the forecast error variance, but it is perfectly possible for the shape of the lag structure to be wave-like, with some higher lags more siginficant than lower lags. It is therefore a potential trade-off between statistical efficiency and economic relevance.

VAR Lag Order Selection Criteria
Endogenous variables: DLGDP DLP
Exogenous variables: C
Date: 04/21/03 Time: 10:52
Sample: 1947:1 2002:4
Included observations: 211

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-1079.034	NA	96.64199	10.24677	10.27854	10.25961
1	-951.1699	252.0922	29.87284	9.072700	9.168014*	9.111228
2	-941.9000	18.10039	28.41771*	9.022749*	9.181605	9.086961*
3	-938.4256	6.718359	28.56051	9.027730	9.250129	9.117628
4	-936.6418	3.415289	29.16843	9.048738	9.334678	9.164321
5	-930.9297	10.82863*	28.70135	9.032509	9.381992	9.173777
6	-930.6199	0.581514	29.72669	9.067487	9.480512	9.234440
7	-927.0622	6.609516	29.85660	9.071680	9.548247	9.264318
8	-924.3163	5.049336	30.22027	9.083567	9.623677	9.301890
9	-920.5025	6.940747	30.28207	9.085332	9.688984	9.329340
10	-915.7497	8.559614	30.07710	9.078196	9.745390	9.347889
11	-914.5581	2.123438	30.90138	9.104816	9.835553	9.400194
12	-909.2234	9.405202	30.52809	9.092165	9.886444	9.413228

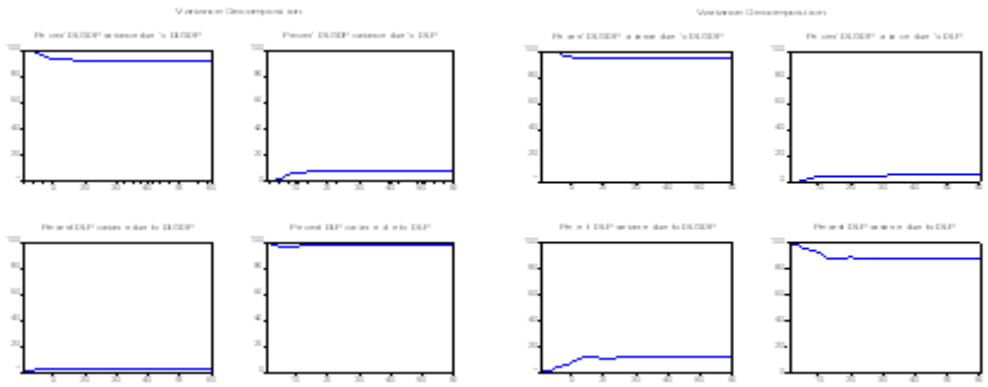
* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

To see if the lag makes much of a difference we look at the impulse response functions for VARs of lag lengths 5 and 12.

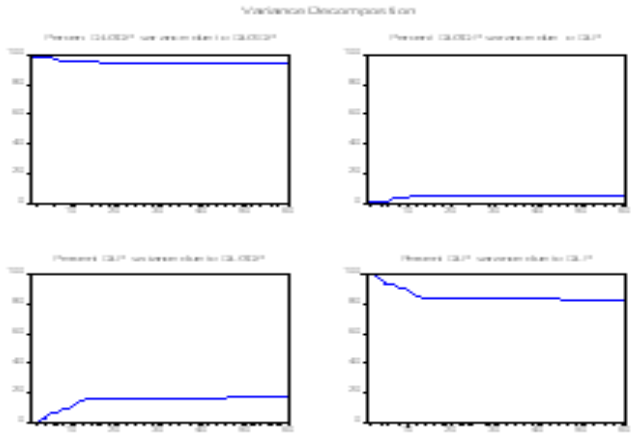


Although the shapes are similar, the VAR with lag length 12 has more action in the IRF than that with lag 5.

The forecast error variance decomposition for the two VARs, based on a Choleski decomposition with Δy_t ordered first, are respectively



Puttting Δp_t first in the Choleski ordering for the VAR(12) gives



The clear message from these variance decompositions is that each variable is almost entirely explained by past own shocks. Growth is almost completely unexplained by inflation, and inflation is little explained by growth. Using a longer lag increases the explanatory power of growth for inflation. This suggests that perhaps the longer lag is preferable.

As far as monetary policy is concerned, controlling inflation by using interest rates to control output seems likely to be rather ineffective in the sense that shocks to output brought about by, for example, unanticipated interest rate changes seem to have only a weak impact of inflation. We have not, of course, established that interest rates do affect output; only that output shocks from whatever cause seem to have a weak effect on inflation.