EC402: Vector Autoregressions (I)

Christian Julliard

Department of Economics and FMG
London School of Economics

- We will drop the simultaneous equations assumption that some variables are endogenous and that other are exogenous, and we will treat all the variables as potentially endogenous.

- This is appealing since in economics, differently from laboratory experiments, everything is endogenous in some sense.

Outline

Outline

The Reduced from VAR

- Let y_t be a $n \times 1$ vector of time series. A reduced form VAR can be written as

$$y_t = \underbrace{c}_{n\times 1} + \underbrace{\Phi_1}_{n\times n} y_{t-1} + \Phi_2 y_{t-2} + ... + \Phi_p y_{t-p} + \varepsilon_t, \quad (1)$$

$$\varepsilon_t \sim iidN\left(0; \underbrace{\Omega}_{n\times n}\right); \; t = -p+1, .., 0, .., T$$

  where c is a vector of constants and the Φ_i are matrixes of coefficients.
- Each variable can potentially depend on its own lags and the lags of all other variables.
- The covariance matrix of the errors is not restricted to be diagonal.

- To write the likelihood we can use the standard time series approach we have already seen (factorization using the conditionals and the marginal distribution) since

$$y_t|y_{t-1},...,y_{-p+1} \sim N\left(c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + ... + \Phi_p y_{t-p}; \Omega\right) \quad (2)$$

Define:

- the $n \times (np+1)$ matrix $\Pi' = [c, \Phi_1, ..., \Phi_p]$ i.e. the $j-th$ row of Π' contains the parameters of the $j-th$ equation.
- the $(np+1) \times 1$ vector $x_t' = \left[1, y_{t-1}', y_{t-2}', ..., y_{t-p}'\right]$.

- We can then rewrite equation (1) as

$$y_t = \Pi' x_t + \varepsilon_t$$

and (2) as

$$y_t|y_{t-1},...,y_{-p+1} \sim N\left(\Pi' x_t; \Omega\right)$$

- Therefore, the conditional pdf of the $t-th$ observation will be given by the multivariate normal

$$p\left(y_t|y_{t-1},...,y_{-p+1}\right) = (2\pi)^{-\frac{n}{2}}\left|\Omega^{-1}\right|^{\frac{1}{2}} \times \exp\left\{-\frac{1}{2}\left(y_t-\Pi' x_t\right)'\Omega^{-1}\left(y_t-\Pi' x_t\right)\right\}.$$

- Treating the first p observation as given, the log likelihood will then be

$$\log L(\Pi,\Omega) = -\frac{Tn}{2}\log 2\pi + \frac{T}{2}\log\left|\Omega^{-1}\right| - \frac{1}{2}\sum_{t=0}^{T}\left(y_t-\Pi' x_t\right)'\Omega^{-1}\left(y_t-\Pi' x_t\right)$$

- Therefore, the MLE will be

$$\underbrace{\hat{\Pi}'_{MLE}}_{n\times(np+1)} = \left[\sum_{t=0}^{T}\left(y_t x_t'\right)\right]\left[\sum_{t=0}^{T}\left(x_t x_t'\right)\right]^{-1}.$$

- Defining with $\hat{\pi}_j$ the $j-th$ row of $\hat{\Pi}'_{MLE}$, we have

$$\hat{\pi}_j = \left[\sum_{t=0}^{T}\left(y_{jt} x_t'\right)\right]\left[\sum_{t=0}^{T}\left(x_t x_t'\right)\right]^{-1}$$

that is, the MLE is simply the OLS estimation equation by equation.

Note: we have already seen the SUR result that if all the equations have the same RHS variables GLS is equivalent to OLS equation by equation.

Warning: if we have some restrictions on the Π coefficients, doing OLS equation by equation is not appropriate – we have in this case to maximize numerically the log likelihood.

- The MLE of the covariance matrix will also have the usual form

$$\begin{aligned} \hat{\Omega}_{MLE} &= \frac{1}{T}\sum_{t=0}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t' \\ &\rightarrow \quad \hat{\sigma}_{ij} = \sum_{t=1}^{T}\hat{\varepsilon}_{it}\hat{\varepsilon}_{jt} \end{aligned}$$

  where $\hat{\varepsilon}$ are the estimated residuals and σ_{ij} is the (i,j) element of Ω.
- Moreover, the usual MLE asymptotic results for the parameters estimate apply.

- It is worth noticing that the likelihood evaluated at its peak has a very simple form

$$\log L\left(\hat{\Pi},\hat{\Omega}\right) = -\frac{Tn}{2}\log 2\pi + \frac{T}{2}\log\left|\hat{\Omega}^{-1}\right| - \frac{1}{2}\sum_{t=0}^{T}\underbrace{\hat{\varepsilon}_t'\hat{\Omega}^{-1}\hat{\varepsilon}_t}_{1\times 1} \tag{3}$$

- Since the *trace* of a scalar is the scalar itself we have that

$$\begin{aligned}\sum_{t=0}^{T}\hat{\varepsilon}_t'\hat{\Omega}^{-1}\hat{\varepsilon}_t &= trace\left[\sum_{t=0}^{T}\hat{\varepsilon}_t'\hat{\Omega}^{-1}\hat{\varepsilon}_t\right] = trace\left[\sum_{t=0}^{T}\hat{\Omega}^{-1}\hat{\varepsilon}_t'\hat{\varepsilon}_t\right] \\ &= trace\left[\hat{\Omega}^{-1}\left(T\hat{\Omega}\right)\right] = trace\left[T\times I_n\right] = Tn\end{aligned}$$

(where I_n is the $n\times n$ identity matrix), the likelihood evaluated at the MLE becomes simply

$$\log L\left(\hat{\Pi},\hat{\Omega}\right) = -\frac{Tn}{2}\log 2\pi + \frac{T}{2}\log\left|\hat{\Omega}^{-1}\right| - \frac{Tn}{2} \tag{4}$$

$\Rightarrow$ Constructing the *LR* test is straightforward.

- Suppose we want to compare a restricted model (indexed by 0) and an unrestricted one (indexed by 1), we then have

$$\begin{aligned} LR &= 2\left[\frac{T}{2}\log\left|\hat{\Omega}_1^{-1}\right| - \frac{T}{2}\log\left|\hat{\Omega}_0^{-1}\right|\right] \\ &= T\left[\log\left(1/\left|\hat{\Omega}_1\right|\right) - \log\left(1/\left|\hat{\Omega}_0\right|\right)\right] \\ &= T\left[\log\left|\hat{\Omega}_0\right| - \log\left|\hat{\Omega}_1\right|\right] \sim \chi^2_{(\#\text{of restrictions})} \end{aligned}$$

- In small sample Sims (1980) suggested to use a slightly different construction of the LR statistic

$$LR := [T - (1 + np_1)]\left[\log\left|\hat{\Omega}_0\right| - \log\left|\hat{\Omega}_1\right|\right] \sim \chi^2_{(\#\text{of restrictions})}$$

where $(1 + np_1)$ is the number of parameters per equation in the unrestricted model.

Example

Suppose we want to choose between two different lag lenghts $p_1 > p_0$. We can then simply proceed as follows:

1. Run OLS equation by equation using in turn p_1 and p_0 lags
2. Construct $\hat{\Omega}_1$ and $\hat{\Omega}_0$ from the OLS residuals
3. Form the LR statistic that will be distributed as $\chi^2_{(n^2(p_1-p_0))}$ (since the difference in the number of lags is $p_1 - p_0$ for each variable in each equation and we have n variables and n equations)

Outline

Akaike and Bayesian Information Criteria

- The Akaike and the Bayesian information criteria (AIC and the BIC) are often used in VAR model selection.
- BIC and AIC are not based on the comparison between a statistic and a distribution. Instead, they provide a way to "rank" alternative specifications and provide a decision rule that, as the sample size goes to infinity, will deliver the right choice with probability one

Note: all the tests we have seen so far imply that even if $T \to \infty$ we will always be making the wrong choice with positive probability (given by the chosen confidence level)

Idea: the best fitting model will be characterized by the sharpest likelihood.

But: a model with more free parameters will generally fit better than a model with a smaller number of parameters.

- BIC and AIC: look at the value of the log likelihood at its peak and introduce a penalty for dimensionality.

Definition (BIC and AIC)

$$BIC := -2 \log L\left(\hat{\Pi}, \hat{\Omega}\right) + d \times \log T$$

$$AIC := -2 \log L\left(\hat{\Pi}, \hat{\Omega}\right) + 2d$$

where d is the number of independent parameters in $\hat{\Pi}$ and $\hat{\Omega}$.

- For both criteria the smaller the better (there is a minus sign in from of the log likelihood).
- The second term is a penalty for the dimensionality of the model.

Note: the BIC has larger penalty term than the AIC, and will therefore tend to favor more parsimonious models.

- Given the simple form taken by the $\log L\left(\hat{\Pi}, \hat{\Omega}\right)$ (see (4)) both statistics are very simple to construct (and are normally computed by econometrics software).

- For example, if we wanted to decide how many lags to include in a VAR, we could compute the BIC, or the AIC, for each of the lag length considered and we would pick the model that delivers the lowest value. This selection criterion (under regularity conditions) would deliver the right choice with probability one as $T \to \infty$.

Outline

Granger Causal Ordering

In modeling relationships among variables in a VAR setting, is helpful to introduce some formal concept of *causality*.

Definition (Granger Causality)

X does not Granger-cause Y ($X \sim GC\ Y$) iff prediction of Y based on the universe of predictors U is not better than prediction based on $U - \{X\}$ i.e. the universe with X omitted.

Example

Consider the reduced from VAR

$$\left\{ I - \begin{bmatrix} B_{11}(L) & B_{12}(L) & B_{13}(L) \\ B_{21}(L) & B_{22}(L) & B_{23}(L) \\ B_{31}(L) & B_{32}(L) & B_{33}(L) \end{bmatrix} \right\} \begin{bmatrix} y_t \\ x_t \\ z_t \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix} \quad (5)$$

where I is the identity matrix of appropriate dimension, the $B_{ij}(L)$ are polynomials in the lag operator and the ε's are error terms.

The universe of predictors consists of the past values y, x and z, and we have that

$$\begin{aligned} x &\sim GC\ y \text{ iff } B_{12} = 0;\ x \sim GC\ z \text{ iff } B_{32} = 0; \\ z &\sim GC\ y \text{ iff } B_{13} = 0;\ z \sim GC\ x \text{ iff } B_{23} = 0; \\ y &\sim GC\ x \text{ iff } B_{21} = 0;\ y \sim GC\ z \text{ iff } B_{31} = 0; \end{aligned}$$

Note that:

1. Granger-causality does not "discover" any true causal structure if we don't have a supporting theory: it is only a necessary condition for a causal relation.
2. Granger-causality is *not transitive*, that is the fact that y *GC* z and that z *GC* x does not imply that y *GC* x. This is not the way we normally think about causality.

In the VAR (5) if $B_{12} = 0$ but all the other coefficients are different from zero, we have that x *GC* z and z *GC* y but $x \sim GC\ y$.

Nevertheless, we can define a transitive relation based on Granger-causality

Definition (Granger Causal Priority)

x is Granger Causal Prior to y (x *GCP* y) in a system like (5) iff it is possible to group all the variables in the system into two blocks, Y_1 and Y_2, such that y is in Y_1 and x is in Y_2, and $Y_1 \sim GC\ Y_2$.

Note: if x *GCP* y, it is also true that $y \sim GC\ x$.

Example

In (5):

$$x \text{ GCP } y \text{ iff either } B_{21} = B_{23} = 0 \text{ or } B_{21} = B_{31} = 0$$

- Testing for *GCP* and $\sim$ *GC* is simple since they both imply linear parameter restrictions.
- So, we can use any of the tests of parameter restrictions seen in the previous lectures.
- For example, we could just estimate the unrestricted and restricted models and then from the *LR* test statistic.

Outline

VAR properties from the Jordan decomposition

We can always rewrite the VAR

$$\underbrace{X_t}_{n\times 1} = \sum_{s=1}^{k} B_s X_{t-s} + \varepsilon_t \tag{6}$$

as

$$Y_t = AY_{t-1} + \eta_t$$

where

$$Y_t = \begin{bmatrix} X_t \\ X_{t-1} \\ \dots \\ X_{t-k+1} \end{bmatrix}; \; A = \begin{bmatrix} B_1 & B_2 & \dots & B_k \\ I_{(k-1)\times n} & & & \underline{0} \end{bmatrix}; \; \eta_t = \begin{bmatrix} \varepsilon_t \\ \underline{0} \end{bmatrix}$$

and A is a square matrix

The Jordan decomposition of A is

$$A = P\Lambda P^{-1}$$

where Λ is diagonal except that it might contain "Jordan blocks" of the form

$$\begin{bmatrix} \lambda & 1 & 0 & \dots & \dots & 0 \\ 0 & \lambda & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 1 & 0 \\ 0 & \dots & \dots & 0 & \lambda & 1 \\ 0 & \dots & \dots & \dots & 0 & \lambda \end{bmatrix}$$

where the λ_i are eigenvalues and P is a matrix with columns given by the eigenvectors of A

Note: this is handy since $A^m = P\Lambda^m P^{-1}$. Therefore, if we want to study the long run behavior of our VAR, we can simply look at the eigenvalues of A.

- If we define $Z_t = P^{-1} Y_t$ we have that

$$\begin{aligned} Z_t &= \Lambda Z_{t-1} + \tilde{\eta}_t \text{ where } \tilde{\eta}_t := P^{-1}\eta_t \\ \rightarrow & \quad z_{i,t} = \Lambda_i z_{i,t-1} + \tilde{\eta}_{i,t} \end{aligned}$$

where i refers to the subsystem corresponding to the Jordan block Λ_i with λ_i on the main diagonal.

$$\therefore z_{i,t} = \Lambda_i^t z_{i,0} + \sum_{s=0}^{t-1} \Lambda_i^s \tilde{\eta}_{i,t-s}$$

- Note that Λ_i^p has λ_i^p on the main diagonal and the $q-th$ diagonal above the main contains:
  - $\frac{p!}{q!(p-q)!}\lambda^{p-q}$ for $q \leq p$
  - 0 for $q > p$

So, since Y is a linear combination of Z

1. If $|\lambda_i| < 1 \; \forall i$, Y (and hence X) is stationary
2. if $\exists i$ s.t. $|\lambda_i| = 1$ and $|\lambda_j| \not> 1 \forall j \neq i$, Y contains components that eventually grow at rate t^m where m is the order of the largest Jordan block with $|\lambda_i| = 1$
3. if $\exists i$ s.t. $|\lambda_i| > 1$, Y contains components that explode at exponential rate
4. if $\exists i$ s.t. λ_i is complex, Y has elements that show a cyclical component

- It is often useful to look at the eigenvectors (columns of P) corresponding to the various types of roots.
  Ex.:consider a VAR containing nominal values in a country with high and variable inflation. We should expect one unstable root to correspond to the price level, and also that the row of P^{-1} corresponding to this root should put positive weight on a set of nominal variables (if the variables are in logs we should expect the same number for this weights)

We can also link these results to the concepts of stationarity and cointegration we have seen in the previous lectures.

Proposition:

In a VAR with n variables, if:

1. there are m unstable λ_i,
2. they are all equal
3. their Jordan blocks are diagonal (this can be relaxed)

then, there are $n - m$ **stationary linear combinations** of X_t

Note: We normally focus on unstable roots exactly equal to 1, but the result above is more general.

Outline

Vector Error Correction Models

As for the single dynamic equation case, a VAR for a $n \times 1$ vector of time series X_t

$$X_t = \sum_{s=1}^{p} \underbrace{B_s}_{n \times n} X_{t-s} + \varepsilon_t$$

can be rewritten in error correction form as

$$\Delta X_t = \sum_{s=1}^{p-1} \underbrace{G_s}_{n \times n} \Delta X_{t-s} + G_0 C X_{t-1} + \varepsilon_t \tag{7}$$

This is the VECM representation, that is a particular VAR.

- This representation is handy because if some of the variables in X_t are $I(1)$ but cointegrated, if we defined with C the $h \times n$ matrix containing the h cointegrating vectors (one for each row), we have that CX_t is stationary, making the all system stationary.

Note:

1. C is not of full rank since with n variables there is a maximum of $n-1$ linearly independent cointegration relationships.
2. C is not unique since if CX_t is stationary, for any $h \times h$ non-zero matrix A, we have that ACX_t is also stationary

- Equation (7) makes clear that a VAR in first differences (obtained setting G_0 equal to zero) is not consistent with a cointegrated system since it would rule out cointegration.
- Nevertheless, a VAR in levels does not have this problem.

- The VECM form is problematic because we generally don't know ex-ante which linear combinations are stationary.
- It is good if economic theory tells us which combinations of variables should be stationary and which should not
- Usually instead researchers claim not to know C and the number of cointegrating relationships (h) in it, and try to estimate it to then write the VECM form.
- The problem is that there is a large set of potential cointegrating relationships, and for each of them there are several possible VECM.
- Moreover, classical model selection is problematic since with unstable roots there is no asymptotic Gaussianity of the parameter estimated.

- The classical approach in modeling a VECM is the following:
  1. Try to estimate C.

Problem: there are many way of doing this (one is OLS one variable at the time as we have seen) and:

- generally give very different results,
- $\hat{C}$ will tend to vary a lot in small sample.

  2. Act as $\hat{C}$ is the "true" one and proceed.