EC402: MEI. Spring Term: The Analysis of Time Series*

Lecture Notes #6: Vector Autoregressions

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.

1. Reduced form VAR

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

$$\begin{aligned} y_t &= \underbrace{c}_{n\times 1} + \underbrace{\Phi_1}_{n\times n} y_{t-1} + \Phi_2 y_{t-2} + ... + \Phi_p y_{t-p} + \varepsilon_t, \\ \varepsilon_t &\sim iidN\left(0; \underbrace{\Omega}_{n\times n}\right);\ t = -p+1, .., 0, .., T \end{aligned} \tag{1.1}$$

where c is a vector of constants and the Φ_i are matrixes of coefficients. In this model each variable can potentially depend on its own lags and the lags of all other variables. Moreover, the covariance matrix of the errors is not restricted to be diagonal.

Conditioning on the first p observations and on the parameters (that we summarize by θ) the joint pdf of the data will be

$$p\left(y_T, y_{T-1}, ..., y_1 | y_0, y_{-1}, ..., y_{-p+1}, \theta\right)$$

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

$$y_t|y_{t-1},...,y_{-p+1} \sim N\left(c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + ... + \Phi_p y_{t-p}; \Omega\right) \tag{1.2}$$

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. Define 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 (1.1) as

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

and (1.2) 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

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

This means that the log-likelihood, treating the first p observation as given, will be

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

implying that[1]

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

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. This is not surprisingly since we have seen the SUR result that if all the equations have the same RHS

[1] See Hamilton, page 293 for a derivation.

variables GLS is equivalent to OLS equation by equation. This also implies that if we have some restrictions on the Π coefficients, this approach is not appropriate – we should in this case maximize numerically the log likelihood.

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

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

where $\hat{\varepsilon}$ are the estimated residuals and σ_{ij} is the (i,j) element of Ω.

Moreover, the usual MLE asymptotic results of the parameters estimate apply.

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

$$\log L\left(\hat{\Pi},\hat{\Omega}\right) = -\frac{Tn}{2}\log 2\pi + \frac{T}{2}\log\left|\hat{\Omega}^{-1}\right| - \frac{1}{2}\sum_{t=0}^{T}\underbrace{\hat{\varepsilon}_t'\hat{\Omega}^{-1}\hat{\varepsilon}_t}_{1\times 1} \tag{1.3}$$

Since the *trace* of a scalar[2] 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

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

[2]Recall that the *trace* of a $n\times n$ matrix A is defined as the sum of the elements along the principal diagonal

$$trace\left(A\right) = a_{11} + a_{22} + ... + a_{nn}.$$

Recall also that if A is $m\times n$ and B is $n\times m$, then

$$trace\left(AB\right) = trace\left(BA\right)$$

and that if A and B are both $n\times n$, then

$$trace\left(A+B\right) = trace\left(A\right) + trace\left(B\right)$$

This implies that the construction of the likelihood ratio test is straightforward. Suppose we want to compare a restricted model (denoted by the index 0) and an unrestricted one (denoted by the index 1), the likelihood ratio test will simply be

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

Example 1. *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. *Then 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)*

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.

1.1. Akaike and Bayesian Information Criteria

The Akaike information criterion, AIC, and the Bayesian information criterion, BIC (also called the Schwartz criterion), are often used in VAR model selection (but can also be used in other settings).

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

(while instead 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 – with probability given by the chosen confidence level).

Both BIC and AIC are based on the same idea: the best fitting model will be characterized by the sharpest likelihood. Therefore, they are both based on the value of the likelihood evaluated at its peak. Nevertheless, since a model with more parameters will generally fit better than a model with a smaller number of parameters, both statistics introduce a penalty for dimensionality.

Definition 1. *Bayesian Information Criterion*

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

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

Definition 2. *Akaike Information Criterion*

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

For both criteria the smaller the better (since there is a minus sign in from of the log likelihood evaluated at its peak). The second term is a penalty for the dimensionality of the model. Note that the BIC has larger penalty term than the AIC, and it will therefore tend to favor more parsimonious models.

Given the simple form taken by the log likelihood at its peak (see (1.3)) 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 $T \to \infty$.

1.2. Granger Causality and Causal Ordering

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

Definition 3. *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 2. *Consider the reduced from VAR*

$$\left\{ I - \begin{bmatrix} B_{11}(L) & B_{12}(L) & B_{13}(L) \\ B_{21}(L) & B_{22}(L) & B_{23}(L) \\ B_{31}(L) & B_{32}(L) & B_{33}(L) \end{bmatrix} \right\} \begin{bmatrix} y_t \\ x_t \\ z_t \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix} \tag{1.4}$$

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.
In this case the universe of predictors consists of the past values y, x and z. In this case 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.

Example 3. *In the VAR (1.4) 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 4. *x is Granger Causal Prior to y ($x\ GCP\ y$) in a system like (1.4) 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 that if x GCP y, it is also true that $y \sim GC$ x.

Example 4. *In (1.4):*

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

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.

1.3. VAR properties from the Jordan decomposition

We can always rewrite a VAR with k lags (where for simplicity I'm disregarding the constant terms)

$$\underbrace{X_t}_{n\times 1} = \sum_{s=1}^{k} B_s X_{t-s} + \varepsilon_t \tag{1.5}$$

as a VAR with only one lag

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

where

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

Note that A is a square matrix, therefore we can write the Jordan decomposition[3]

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

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

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

[3]See the Appendix for a formal definition.

where the λ_i are eigenvalues and P is a matrix with columns given by the eigenvectors 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. We can solve this last equation backward obtaining

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

Note that Λ_i^p has λ_i^p on the main diagonal and the $q-th$ diagonal above the main contains:

1. $\frac{p!}{q!(p-q)!}\lambda^{p-q}$ for $q \leq p$
2. 0 for $q > p$

Since Y_t is a linear combination of Z_t ($Y_t = PZ_t$), we can state the following results:

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 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 (in the matrix P) corresponding to the various types of roots.

Example 5. *Consider a VAR containing nominal values in a country with high and variable inflation. We should in this case expect one unstable root to correspond to the price level. Moreover, we would expect that the row of P^{-1} corresponding to this root should put positive weights on a set of nominal variables (if the variables are in logs we should expect the same number for this weights).*

We can use the results from the Jordan decomposition an link them to the concepts of Cointegration and Stationarity studied in previous lectures.

Proposition 1. *If i) there are m unstable λ_i, ii) they are all equal and iii) their Jordan blocks are diagonal,*[4] *there are $n-m$* ***stationary*** *linear combinations of X_t.*

1.4. 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

$$\Delta X_t = \sum_{s=1}^{p-1} \underbrace{G_s}_{n \times n} \Delta X_{t-s} + \underbrace{G_0}_{n \times h} \underbrace{C}_{h \times n} X_{t-1} + \varepsilon_t \tag{1.6}$$

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 matrix containing the h linearly independent cointegrating relationships among variables (one cointegration vector for each row) we have that CX_{t-1} is stationary, making the all system stationary.

Note that C is not of full rank since with n variables there is a maximum of $n-1$ linearly independent cointegrating relations (since there is a maximum of $n-1$ common trends), that is $h < n$. Moreover, C is not unique since if CX_t is stationary, for any non zero $h \times h$ matrix A we also have that ACX_t is stationary.

Equation (1.6) makes also 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

[4] This last requirement can be relaxed.

variables should be stationary and which should not. Usually instead researchers claim not to know C and the number of cointegrating relationship 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 relationship, 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.[5]

The classical approach 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 i) generally give very different results, ii) $\hat{C}$ will tend to vary a lot in small sample.

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

2. Structural VAR (S-VAR)

Let X_t be a $n \times 1$ vector of time series. A Structural VAR (S-VAR) takes the form

$$\Gamma_0 X_t + \Gamma_1 X_{t-1} + ...\Gamma_p X_{t-p} = \underbrace{c}_{n\times 1} + \underbrace{\varepsilon_t}_{n\times 1} \text{ where } \varepsilon_t \sim N(0, \Sigma) \tag{2.1}$$

where Σ is of dimension $n \times n$ as well as each Γ_i matrix, n is the number of variables in the system, c is a $n \times 1$ vector of constants and ε_t span the space of *innovations* to X_t.

This structure implies that each variables in the system can potentially depend on past and current values of all the other variables.

Two commonly used normalizations are: $\Sigma = I$ (the identity matrix) or each variable has coefficient 1 in one of the $\Gamma(L)$. We will use the former in what follows.

We also assume Γ_0 is full rank i.e. the system can be solved to determine X_t from past X and ε (the system is "complete"), that is we can rewrite the system in "reduced form"

$$X_t = \gamma + B(L) X_{t-1} + v_t \tag{2.2}$$

[5]This is a problem only of the frequentist approach, since the standard Bayesian asymptotic Gaussian approximation of the likelihood holds even for unit roots.

where $v_t = \Gamma_0^{-1}\varepsilon_t$, $\gamma = \Gamma_0^{-1}c$ and $B(L) = -\Gamma_0^{-1}(\Gamma_1 + \Gamma_2 L.... + \Gamma_p L^{p-1})$. We can estimate equation (2.2) as discussed in Section (1).

2.1. Identification

As in the simultaneous equation case, a key question is whether we can recover the parameters of the structural form (2.1) from the parameters of the reduced form (2.2).

Let's consider the normalization $\Sigma = I$ (we already discussed the other alternative normalization in the case of simultaneous equations).

The reduced form gives in γ and $B(L)$ as many parameters as in c and $\Gamma_1, ...\Gamma_p$. Moreover, we have that

$$v_t \sim N\left(0; \Gamma_0^{-1}\left(\Gamma_0^{-1}\right)'\right)$$

so we could hope to recover Γ_0 from the covariance matrix of v_t. The problem is that there are many $n \times n$ matrixes G such that $GG' = \Gamma_0^{-1}\left(\Gamma_0^{-1}\right)'$. This is due to the fact that there are $(n+1)n/2$ free elements in $\Gamma_0^{-1}\left(\Gamma_0^{-1}\right)'$ while Γ_0 has n^2 free elements. This means that if we want to find the structural parameters we need at least $(n-1)n/2$ restrictions.

The most commonly used approach to obtain identification is to impose restrictions in the Γ_0 matrix alone. There are two good reasons why this is appealing.

First, these restrictions have a natural interpretation as assumptions about delays in reactions of particular variables.

Example 6. *When dealing with quarterly data, it might be natural to assume that, in setting the Federal funds rate, the Federal Reserve Bank reacts contemporaneously to inflation, but that the level of inflation in the economy is not influenced immediately by the actions of the central bank. In writing a S-VAR for the interest rate, i, and inflation, π, this assumption would be formulated as*

$$\Gamma_0 \begin{bmatrix} i_t \\ \pi_t \end{bmatrix} + \Gamma_1 \begin{bmatrix} i_{t-1} \\ \pi_{t-1} \end{bmatrix} + ...\Gamma_p \begin{bmatrix} i_{t-p} \\ \pi_{t-p} \end{bmatrix} = c + \varepsilon_t$$

$$\text{with } \Gamma_0 = \begin{bmatrix} \times & \times \\ 0 & \times \end{bmatrix}$$

where $\times$ denotes non-zero elements.

Second, the structural estimation can be performed in a relatively simple way by first estimating γ and $B(L)$ performing OLS equation by equation in the reduced form, and then maximizing

$$-\frac{Tn}{2}\log 2\pi+\frac{T}{2}\log\left|\left(\Gamma_0^{-1}\left(\Gamma_0^{-1}\right)'\right)^{-1}\right|-\frac{1}{2}trace\left[\left(\Gamma_0^{-1}\left(\Gamma_0^{-1}\right)'\right)^{-1}\left(\sum_{t=0}^{T}\hat{v}_t'\hat{v}_t\right)\right]$$

with respect to Γ_0, where $\hat{v}_t$ are the reduced form OLS residuals.[6] This last step is straightforward if we have exactly $(n-1)n/2$ restrictions, but it requires numerical optimization if we have more than $(n-1)n/2$ restrictions (i.e. if the system is over-identified).

Note that imposing at least $(n-1)n/2$ restrictions in the Γ_0 matrix is only a necessary condition, but it is not sufficient since we also need the restrictions to be linearly independent otherwise Γ_0 wouldn't be invertible.

Example 7. *In a 2 variables system we need at least* $(n-1)n/2=(2-1)2/2=1$ *zero restrictions. Examples that works are* Γ_0 *of the form*

$$\begin{bmatrix}\times & 0\\ \times & \times\end{bmatrix} \text{ or equivalently } \begin{bmatrix}0 & \times\\ \times & \times\end{bmatrix}.$$

A useless Γ_0 *is*

$$\begin{bmatrix}\times & 0\\ \times & 0\end{bmatrix}$$

since it is not invertible.

In a 3 variables system we need at least $(n-1)n/2=(3-1)3/2=2$ *zero restrictions. An exactly identified* Γ_0 *is then of the form*

$$\begin{bmatrix}\times & \times & \times\\ 0 & \times & \times\\ 0 & 0 & \times\end{bmatrix}.$$

An interesting one is

$$\begin{bmatrix}\times & 0 & \times\\ 0 & \times & \times\\ 0 & \times & \times\end{bmatrix}.$$

In this case we have enough zeros but nevertheless we don't have identification since it would not be invertible (the last two rows are not linearly independent). In this case, even

[6] Note that this expression is simply the likelihood maximized with respect to γ and $B(L)$.

though we can identify the first equation with respect to the other two, we cannot identify the last two equations and the system is not fully identified.

2.2. Impulse-Response Functions

Given the S-VAR (where for simplicity I'm disregarding the vector of constant terms)

$$\Gamma_0 X_t = A(L) X_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \sim N(0, I) \tag{2.3}$$

we might be interest in how a shock in one of the equations influences the other variables in the system. In Example (6) we might want to know how a shock to the current interest rate would affect inflation in the future.

The system (2.3) can be rewritten as

$$X_t = \Gamma_0^{-1} A(L) X_{t-1} + \Gamma_0^{-1} \varepsilon_t$$

and this implies that

$$\begin{aligned} X_{t+1} &= \Gamma_0^{-1} A(L) X_t + \Gamma_0^{-1} \varepsilon_{t+1} \\ &= \left[\Gamma_0^{-1} A(L)\right]^2 X_{t-1} + \Gamma_0^{-1} A(L) \Gamma_0^{-1} \varepsilon_t + \Gamma_0^{-1} \varepsilon_{t+1} \\ X_{t+2} &= \Gamma_0^{-1} A(L) X_{t+1} + \Gamma_0^{-1} \varepsilon_{t+2} \\ &= \left[\Gamma_0^{-1} A(L)\right]^3 X_{t-1} + \left[\Gamma_0^{-1} A(L)\right]^2 \Gamma_0^{-1} \varepsilon_t + \Gamma_0^{-1} A(L) \Gamma_0^{-1} \varepsilon_{t+1} + \Gamma_0^{-1} \varepsilon_{t+2} \end{aligned}$$

We can therefore define the Impulse Response Function (IRF) of the $j-th$ variable in X to a shock in the $i-th$ equation as

$$\frac{\partial E_t [X_{j,t+S}]}{\partial \varepsilon_{i,t}} = \left\{ \left[\Gamma_0^{-1} A(L)\right]^S \Gamma_0^{-1} \right\}_{ji} \tag{2.4}$$

where $\{\}_{ji}$ denotes the (j,i) element.

The IRF's are a very useful data summary since they concisely report the link between variables over time. For example we would like to know what will be the effect of a monetary policy shock on GDP and inflation in the future.

Note that we can obtain correct IRF's for some of the shocks even if the system is not fully identified, that is even if the system is only *partially identified*.

Example 8. *Suppose we want to write a S-VAR for inflation, π, output gap, x and the interest rate, i. Moreover, suppose we believe that the central bank reacts contemporaneously to news about inflation and output gap (in a Taylor rule fashion), but that both inflation and output gap reacts with a lag to monetary policy shocks. These assumptions will deliver the S-VAR model*

$$\underbrace{\begin{bmatrix} \times & \times & 0 \\ \times & \times & 0 \\ \times & \times & \times \end{bmatrix}}_{\Gamma_0} \begin{bmatrix} \pi_t \\ x_t \\ i_t \end{bmatrix} = \underbrace{A(L)}_{3\times 3} \begin{bmatrix} \pi_{t-1} \\ x_{t-1} \\ i_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t^{\pi} \\ \varepsilon_t^{x} \\ \varepsilon_t^{i} \end{bmatrix};\ \varepsilon_t \sim N(0, I).$$

Since to reach identification we need $n(n-1)/2 = 3$ zeros, this Γ_0 will not deliver identification unless we add a zero in one of the first two equations. But adding a zero in the first two equation would imply that either inflation reacts with a lag to output gap, or that output gap react with a lag to inflation, and any of these assumption might be hard to justify.

Nevertheless, it is interesting to notice that the third row of Γ_0 is identified with respect to the first two ones. That is, if we considered the model as being composed by two blocks, the first one containing the first two rows and the second one containing the third one, we have that the two blocks are identified with respect to each other. This implies that the impact of a monetary policy shock (ε^i) on the upper block is well identified. So, if we are only interested in the IRF's of an ε^i shock, we can simply add an arbitrary zero restriction in one of the first two rows of Γ_0 and equation (2.4) will give the appropriate answer for the ε^i shocks (but not for the others).

Partial identification schemes are often used in applied research. For example, Christiano, Eichenbaum, Evans (1999) study the following S-VAR model of the US economy

$$\Gamma_0 \begin{bmatrix} X_{1t} \\ i_t \\ X_{2t} \end{bmatrix} = \Gamma(L) \begin{bmatrix} X_{1t-1} \\ i_{t-1} \\ X_{2t-1} \end{bmatrix} + \varepsilon_t \sim iidN(\mathbf{0}, I)$$

where X_{1t} and X_{2t} are vectors of dimensions $n_1 \times 1$ and $n_2 \times 1$ (with $n_1, n_2 > 1$), i_t is the Federal Funds interest rate, Γ_0 is a square matrix of appropriate dimensions, $\Gamma(L)$ is a matrix with elements given by polynomials of order k in the lag operator, $\mathbf{0}$ is a column

vector of zeros and I is the identity matrix. In order to identify the effect of monetary policy shocks, i.e. shocks to the Federal Funds interest rate, these authors assume that: 1) the interest rate, i_t, reacts contemporaneously to the X_{1t} variables but with a lag to the X_{2t} variables, 2) the X_{2t} variables react contemporaneously to all the variables in the systems and 3) the X_{1t} variables react with a lag to all the other variables in the system. That is, they assume the following form of the Γ_0 matrix:

$$\Gamma_0 = \begin{bmatrix} \underbrace{\gamma_{11}}_{n_1 \times n_1} & \underbrace{0}_{n_1 \times 1} & \underbrace{0}_{n_1 \times n_2} \\ \underbrace{\gamma_{21}}_{1 \times n_1} & \underbrace{\gamma_{22}}_{1 \times 1} & \underbrace{0}_{1 \times n_2} \\ \underbrace{\gamma_{31}}_{n_2 \times n_1} & \underbrace{\gamma_{32}}_{n_2 \times 1} & \underbrace{\gamma_{33}}_{n_2 \times n_2} \end{bmatrix}. \tag{2.5}$$

Obviously, this model is not fully identified. Nevertheless, the row corresponding to the interest rate is linearly independent form all the other rows in the system. That is, shocks to this equation are correctly identified. So, if we are only interested in the IRF's of an monetary policy shock, we can simply add arbitrary - linearly independent - zero restrictions in the other two blocks of Γ_0 and equation (2.4) will deliver appropriate impulse responses for this shock (but not for the others).

2.2.1. The Cholesky Decomposition

A very popular approach to identification of the IRF's from a reduced form VAR of the form

$$X_t = B(L) X_{t-1} + v_t \text{ where } v_t \sim N(0, \Omega) \tag{2.6}$$

is the *Cholesky decomposition.* This approach is based on the simple fact that for any real symmetric positive definite matrix Ω there exist a unique lower triangular matrix A with 1s along the main diagonal and a unique diagonal matrix D with positive entries along the main diagonal such that

$$\Omega = ADA' = AD^{1/2}D^{1/2}A' = PP' \text{ where } P := AD^{1/2}$$

Note that P is also lower triangular.

Using A we can construct an $n \times 1$ vector $u_t := A^{-1}v_t$. Note that the covariance matrix of u is diagonal since

$$\begin{aligned} E\left(u_t u_t'\right) &= \left(A^{-1}\right) E\left(v_t v_t'\right) \left(A^{-1}\right)' = \left(A^{-1}\right) \Omega \left(A'\right)^{-1} \\ &= \left(A^{-1}\right) A D^{1/2} D^{1/2} A' \left(A'\right)^{-1} = D \end{aligned}$$

that is, the elements of u are *orthogonalized errors* with variance given by D.

Note that similarly $\eta_t := P^{-1}v_t = D^{-1/2}u_t$ is also a vector of orthogonalized errors.

Using these observations, researchers often reports IRF's given by either[7]

$$\frac{\partial E_t\left[X_{j,t+T}\right]}{\partial u_{i,t}} \text{ or } \frac{\partial E_t\left[X_{j,t+T}\right]}{\partial \eta_{i,t}} \tag{2.7}$$

The Cholesky decomposition approach in not a magic wand that delivers meaningful identification of the IRF's. But instead should be thought of as one particular set of restrictions on the Γ_0^{-1} matrix of the structural form, and it will make sense iff this set of restrictions does. To see this note that if we think of the Cholesky decomposition as identifying the effect of the "true" structural shocks – that are assumed to be orthogonal to each other – we should have

$$\Gamma_0^{-1} = AD^{1/2} := \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ a_{21} & 1 & 0 & 0 & 0 \\ a_{31} & a_{32} & 1 & 0 & 0 \\ ... & ... & ... & ... & ... \\ a_{n1} & a_{n2} & a_{n3} & ... & 1 \end{bmatrix} D^{1/2} \tag{2.8}$$

Therefore, the Choleski orthogonalization approach will make economic sense iff this set of zero restrictions are economically meaningful.

In particular, this approach imposes an *order* among the variables. Equation (2.8) implies that the first variable in the system will respond contemporaneously only to its own structural shock, the second variable in the system will respond contemporaneously to its own structural shock and to the structural shock of the first variable, and so on. If such

[7] This two are the same up to a scale factor equal to the (i,i) element of $D^{1/2}$, $\sqrt{d_{ii}}$ that is

$$\frac{\partial E_t\left[X_{j,t+T}\right]}{\partial \eta_{i,t}} \equiv \sqrt{d_{ii}} \frac{\partial E_t\left[X_{j,t+T}\right]}{\partial u_{i,t}}$$

a natural ordering among the variables applies, then this approach is appropriate, but if it doesn't the IRF's constructed in this fashion will be misleading.

3. S-VAR's, SEM's and the Lucas' Critiques

[This section is provided for your information only and will not be examined]

Lucas (1976) observed that simultaneous equation models (SEM's) that assume that private agents' expectations are *fixed* linear function of past data, are likely to be inappropriate to evaluate the impact of systematic changes in monetary policy.

This is due to the fact that a policy change would be likely to change the agents' optimal forecasting rule, and thus change the dynamic of the private sector behavior that the SEM's assumes to be constant.

Given the similarities between S-VAR's and SEM's, the use of SVAR's to analyze the effects of variations in monetary policy is sometimes taken to be subject to the Lucas Critique.

However, the SVAR's, unlike the old SEM's, do not contain fixed-coefficient expectation rules. Moreover, VAR's are best thought of as a linear approximations to the behavior of the private sector and monetary authorities (Sims(1987), Leeper and Zha(2001)), and the behavior they model implicitly includes dynamics arising from revision in the forecasting rules (as well as other sources of dynamics). That is, they are meant to capture some local linear approximation to the actual nonlinear behavioral rules. As a consequence, a SVAR may do a good job in projecting the impact of monetary policy shocks as long as the model's nonlinearity is not too severe. This implies that if the model appears to fit historical data well and shows little sign of nonlinearity in the sample period, then policy changes that produce policy equation shocks with patterns similar to what has been observed in the past will probably be projected accurately by the model.

4. Appendix

Here we recall some useful results and definitions from linear algebra.

Theorem 1 (Cofactor Expansion). *The **cofactor expansion** of the determinant of a $N \times N$ matrix A is*

$$\det(A) = \sum_{n=1}^{N} a_{in} A_{in} = \sum_{n=1}^{N} a_{jn} A_{jn}$$

for any row i and column j, where A_{ij} denotes the (i,j)th cofactor of the matrix A. [8]

Definition 5 (Characteristic Equation). *Given the $N \times N$ real matrix A, the determinantal equation*

$$\det(A - \lambda \cdot I) = 0$$

*is called the **characteristic equation** of A.*

The characteristic function is a polynomial equation of degree N. Therefore, there are N (potentially complex) roots of the characteristic equation and complex roots occur in conjugate pairs. Let's denote the roots by $\lambda_1, ..., \lambda_N$. According to the cofactor expansion we can write

$$\det(A - \lambda \cdot I) = \prod_{n=1}^{N} (\lambda - \lambda_n).$$

(since the coefficient on λ must be one). The roots are not necessarily distinct. If there are only $K \leq N$ distinct roots, λ_k^* $(k = 1, ..., K)$ and denote the multiplicity of the k-th distinct root by m_k, then

$$\det(A - \lambda \cdot I) = \prod_{k=1}^{K} (\lambda - \lambda_k^*)^{m_k}.$$

Definition 6 (Eigenvalue). *A root λ of the characteristic equation of A is an **eigenvalue** of A.*

Definition 7 (Eigenvector). *A vector $x \in C^N$ for which there is a scalar λ such that $Ax = \lambda \cdot x$ is an **eigenvector** of A.*

[8] See the Appendix of Lecture notes #5 if you don't rember the definition of cofactor.

Definition 8 (Block Diagonal Matrix). *A block diagonal matrix has blocks along the main diagonal, and zeros elsewhere.*

Definition 9 (Upper Bidiagonal Matrix). *A matrix B is* ***upper bidiagonal*** *if its (i, j) element is equal to zero unless $i = j$ or $i = j - 1$, for all i and j.*

Theorem 2 (Jordan Decomposition). *For any $N \times N$ matrix A there exists P such that $A = P\Lambda P^{-1}$, where*

1. *Λ is an upper bidiagonal matrix consisting of the eigenvalues of A repeated according to their multiplicities.*

2. *Λ is block-diagonal and each block has one of A's eigenvalues repeated along its main diagonal, 1's along the diagonal above it and 0's elsewhere (that is, Λ has Jordan blocks on the main diagonal).*

3. *P is a matrix with columns given by the eigenvectors of A.*