

Vector Autoregressions II

Empirical Macroeconomics - Lect 2

Dr. Ana Beatriz Galvao

Queen Mary University of London

January 2012



- A VAR(p) model of the $m \times 1$ vector of time series $y_t = (y_{1t}, y_{2t}, \dots, y_{mt})$ with autoregressive order p :

$$y_t = c + A_1 y_{t-1} + \dots + A_p y_{t-p} + \varepsilon_t$$

where A_i are $m \times m$ coefficients matrices and c is a $m \times 1$ of intercepts.

ε_t is a $m \times 1$ vector of disturbances that have the following properties:

$$\begin{aligned} E(\varepsilon_t) &= 0 \text{ (mean zero);} \\ E(\varepsilon_t \varepsilon_t') &= \Sigma_\varepsilon \text{ (full variance-covariance matrix);} \\ E(\varepsilon_t \varepsilon_s') &= 0 \text{ for } s \neq t \text{ (no serial correlation).} \end{aligned}$$



Estimation of Vector Autoregressions I

- Write the VAR(p) as:

$$y_t = Bx_t + \varepsilon_t$$

where

$$B = (c, A_1, \dots, A_p); (m \times (mp + 1))$$

$$x_t = \begin{bmatrix} 1 \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{bmatrix}; ((mp + 1) \times 1).$$



Estimation of Vector Autoregressions II

- Because the disturbances are normal distributed, the conditional density is multivariate normal distributed:

$$y_t | y_{t-1}, \dots, y_{t-p} \sim N(Bx_t, \Sigma_\varepsilon),$$

and the conditional density of the t^{th} observation is:

$$= (2\pi)^{-(m/2)} |\Sigma_\varepsilon^{-1}|^{1/2} \exp \left[(-1/2) (y_t - Bx_t)' \Sigma_\varepsilon^{-1} (y_t - Bx_t) \right].$$

- The likelihood function is the product of each one of these densities for $t = 1, \dots, T$. The log-likelihood function is the sum of the log of all these densities. The log-likelihood is:

$$\begin{aligned} l(B, \Sigma_\varepsilon) &= -(Tm/2) \log(2\pi) + (T/2) \log |\Sigma_\varepsilon^{-1}| \\ &\quad - (1/2) \sum_{t=1}^T \left[(y_t - Bx_t)' \Sigma_\varepsilon^{-1} (y_t - Bx_t) \right]. \end{aligned}$$



Estimation of Vector Autoregressions III

- The value of B that maximise the log-likelihood is the ML estimator of the VAR coefficients:

$$\hat{B} = \left[\sum_{t=1}^T y_t x_t' \right] \left[\sum_{t=1}^T x_t x_t' \right]^{-1}.$$

- This means that the ML estimator of the VAR coefficients is equivalent to the OLS estimator of y_{jt} on x_t , that is, it is equivalent to apply OLS for each equation of the VAR separately.
- The OLS estimator for each separate equation is also equivalent to the system (multivariate) estimator. This is so because the VAR system is a Seemly Unrelated Regression system (SURE) with the same regressors for each equation in the system.



Estimation of Vector Autoregressions IV

- The ML estimator for the variance is:

$$\hat{\Sigma}_\varepsilon = (1/T) \sum \hat{\varepsilon}_t \hat{\varepsilon}_t'$$

where

$$\hat{\varepsilon}_t \equiv y_t - \hat{B}x_t.$$

- The ML estimator of the variance is consistent, but it is biased in small samples, so it is common to use the variance estimator adjusted by the number of degrees of freedom:

$$\tilde{\Sigma}_\varepsilon = \frac{T}{T - mp - 1} \hat{\Sigma}_\varepsilon.$$



Inference I

- It is possible to show that the (asymptotic) distribution of the coefficients of the j^{th} equation of the VAR is

$$\hat{B}_j \approx N \left(B_j, \hat{\sigma}_j \left[\sum_{t=1}^T x_t x_t' \right]^{-1} \right)$$

where $\hat{\sigma}_j = (1/(T - mp - 1)) \sum_{t=1}^T \varepsilon_{jt}^2$, that is, the coefficients' variance can be computed using equation-by-equation OLS estimation.

- Because the coefficients are asymptotically normal, significance tests for each coefficient can be applied by comparing the t-statistic with the normal distribution.



Inference II

- The *Wald statistics* can be also employed to test hypothesis that impose restrictions on the coefficients. Wald statistics are normally compared with the chi-squared distribution, but a F version of the Wald statistic is preferable when the sample is small. Recall that Wald statistics require only the estimation of the model under the alternative hypothesis.
- *Likelihood ratio statistics* (LR) compare the value of the likelihood function under the null and the alternative hypothesis. The computation of LR statistics requires the estimation of both specifications under the null and alternative. LR statistics are normally employed to select the number of lags in a VAR.



Granger Causality I

- A time series variable y_2 fails to Granger-cause y_1 if the mean squared error of a forecast ($s \geq 1$) of y_{1t+s} based on $(y_{1t}, y_{1t-1}, \dots)$ is the same as the mean forecast error of y_{1t+s} based on both $(y_{1t}, y_{1t-1}, \dots)$ and $(y_{2t}, y_{2t-1}, \dots)$ for all $s \geq 1$.
- Recall that the disturbances of the VAR are measures of forecast errors under assumption that the VAR is the true data generating process.
- If it is true that y_2 fails to Granger-cause y_1 , then a VAR(p) ($m = 2$) with the Granger-non-causality restriction imposed is:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} a_{1,11} & 0 \\ a_{1,21} & a_{1,22} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \end{bmatrix} + \dots + \begin{bmatrix} a_{p,11} & 0 \\ a_{p,21} & a_{p,22} \end{bmatrix} \begin{bmatrix} y_{1t-p} \\ y_{2t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↺

Granger Causality II

- We can show that, in the restricted specification, y_2 does not contribute to the forecasting of y_1 for all horizons from $s \geq 1$.
- The null hypothesis of the test for non-Granger-causality of y_{2t} on y_{1t} is:

$$H_0 : a_{1,12} = a_{2,12} = \dots = a_{p,12} = 0.$$

- A Wald statistic can be used to test this hypothesis.
- Note that to identify the direction of causality between y_2 and y_1 , it is also useful to test whether y_1 fails to Granger-cause y_2 . The causality may be *bidirectional*.

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↺

Granger Causality III

- The Granger-causality described above is based on pairs of variables. However, this may not be adequate, because y_2 may cause y_1 because a third variable, say y_3 , helps predict both y_2 and y_1 . This explain why it is recommended to test for Granger-causality in VAR models with a reasonable number of variables able to explain y_2 and y_1 .

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↺

Stock and Watson (2001) Tests

- values are p-values of the Wald test.

A. Granger-Causality Tests			
Dependent Variable in Regression			
Regressor	π	u	R
π	0.00	0.31	0.00
u	0.02	0.00	0.00
R	0.27	0.01	0.00

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↺

Selection of the VAR order p I

- Up to now the autoregressive order is assumed to be known. In practice, we need to choose one. There are two popular methods to choose the lag order.
- The first one is to use a sequence of LR tests. For example, start with a VAR(4), and then test whether a restricted VAR(3) is not rejected. Then if a VAR(3) is not rejected, go to a VAR(2) (See Lutkepohl, p. 143). The problem of this approach is that the sequential structure affects the size of the test. This means that a test with a 5% significance may have a size (type I error) that is not equal to 5%, it could be, say, 10%, just because it is linked to another test made previously.



Selection of the VAR order p II

- To avoid figuring out which one is the right p-value of the test (Eviews p-values do not take into account this problem), an option is to use information criteria. Information criteria are designed to *consistently* find the model that fits better the data from a group of models: VAR(1), VAR(2), VAR(3), VAR(4).
- There are three different criteria to help use to choose p . Each one may choose different models. They differ by the *penalisation* from the inclusion of additional parameters (recall a VAR(4) has m^2 more parameters than a VAR(3)):

$$\begin{aligned}AIC &= \ln |\hat{\Sigma}_\varepsilon(p)| + \frac{2(pm^2)}{T} \\HQ &= \ln |\hat{\Sigma}_\varepsilon(p)| + \frac{2 \ln \ln T(pm^2)}{T} \\SIC(BIC) &= \ln |\hat{\Sigma}_\varepsilon(p)| + \frac{2 \ln T(pm^2)}{T}.\end{aligned}$$



Selection of the VAR order p III

- The penalisation is such that $AIC < HQ < SIC$. This implies that the SIC (Schwarz IC or Bayesian IC) generally chooses models with a smaller p while AIC (Akaike) chooses models with a higher order p .
- For example, in a empirical application, the autoregressive order that minimises SIC is $p = 2$ while that one that minimises AIC is $p = 4$. Which one to choose? For short samples of quarterly data (say 100), it is normally better to assume a small p because the estimates of the coefficients may start getting weird (creating explosive behaviour) with a large p . Otherwise, it may be a good idea to keep $p = 4$ since it is more likely that the disturbances will have no serial correlation (or no remaining information) with a large p . Finally, if using the model for forecasting, economists' wisdom suggests that it is better to use SIC.



Standard errors / confidence intervals for impulse response functions I

- Recall that impulse responses functions are obtained using coefficients of the MA representation of the VAR.
- The Φ_i coefficients are nonlinear transformations of the estimated coefficients: recall $\Phi_i = \sum_{j=1}^i \Phi_{i-j} A_j$.
- The source of uncertainty is A_i , so the gradient (derivatives) with respect to A_i is computed before the calculation of standard errors of the impulse response function. Because the coefficients A_i are asymptotic normal, the distribution of the impulse responses is also normal.



Standard errors/confidence intervals for impulse response functions II

- Assume that:

$$\alpha = \text{vec}(A_1, A_2, \dots, A_p).$$

$$\alpha \approx N(\alpha, \Sigma_{\hat{\alpha}})$$

Then it is possible to show that:

$$\text{vec}(\hat{\Phi}_i) \approx N(\Phi_i, G_i \Sigma_{\hat{\alpha}} G_i')$$

$$G_i = \frac{\delta \text{vec}(\Phi_i)}{\delta \alpha'}$$

- Researchers have shown that the asymptotic confidence intervals of impulse responses may be a bad representation of small sample confidence intervals, especially when time series are very persistent.



Standard errors/confidence intervals for impulse response functions III

- Alternative techniques are the use of monte carlo and bootstrap. These are techniques that use simulation to compute confidence intervals instead of the analytical solution above.
- In the case of monte carlo, it requires the assumption that the residuals are well approximated by a multivariate normal distribution. Evidence of outliers, heteroscedasticity and serial correlation in the residuals suggest that the use of monte carlo methods may be inadequate.
- The bootstrap alternative can be applied when the residuals are not well behaved. A popular bootstrap approach is the one developed by Kilian (1998).



Standard errors/confidence intervals for impulse response functions IV

- Eviews provides both the analytic and the monte carlo alternative to compute confidence intervals for impulse responses.



Bayesian approach to VARs

Advantages argued by Litterman (1986):

- There is no problem in increasing the number of variables in the VAR even if the sample size is short. The Bayesian approach solves the curse of dimensionality by adding "prior information that accurately reflects what is known about the likely value of their coefficients".
- The Bayesian approach may improve accuracy of forecasts by combining useful information "about the future from a wide spectrum of economic data". The information is weighted via coefficient estimates that combine information on the *prior* with evidence from data.
- Instead of choosing the autoregressive order of a VAR, use a large order, and then use a *prior* distribution on the coefficients of longer lags close to zero.



Bayesian Concepts I

- Two random variables: A and B . Rules of probability imply:

$$p(A, B) = p(A|B) p(B)$$

joint probability (conditional prob.)(marginal prob.)

- The Bayes rule is used to learn something about B by using information on A :

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

- In econometrics, we want to learn μ - vector of unknown parameter – conditional on data X :

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)}$$



Bayesian Concepts II

- Because we only want to learn about the conditional distribution of the parameter given the data:

$$p(\mu|X) \propto p(X|\mu) p(\mu)$$

posterior density likelihood function/prior density

- The prior density $p(\mu)$ contains any non-data information available about μ . It summarises what we know about μ before seeing the data.
- The likelihood function $p(X|\mu)$ is the density of the data conditional on the parameters of the model. It is also referred to as the data generating process.
- The posterior density $p(\mu|X)$ summarises all we know about μ after seeing the data. The posterior is a result of an updating rule that combines both data and non-data information.



Bayesian estimation of Vector Autoregressions I

- Recall that:

$$B = (c, A_1, \dots, A_p); (m \times (mp + 1))$$

$$x_t = \begin{bmatrix} 1 \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{bmatrix}; ((mp + 1) \times 1).$$

- Now define

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{bmatrix}; (T \times (mp + 1));$$



Bayesian estimation of Vector Autoregressions II

$$y = \begin{bmatrix} (y_{11} \dots y_{1T})' \\ \vdots \\ (y_{m1} \dots y_{mT})' \end{bmatrix}; (Tm \times 1)$$

and $\alpha = \text{vec}(B)$, that is, an $m(mp + 1) \times 1$ vector.

- The VAR(p) can be written as:

$$y = (I_m \otimes X)\alpha + \varepsilon,$$

where

$$\varepsilon \sim N(0, \Sigma \otimes I_m)$$

- This means that all data are in y , and the parameters are in α and Σ (because there is no ambiguity here $\Sigma = \Sigma_\varepsilon$).



Bayesian estimation of Vector Autoregressions III

- The likelihood of the VAR model is $L(\alpha, \Sigma)$; and it can be decomposed in two parts:

$$\begin{aligned} L(\alpha, \Sigma) &\propto N(\alpha | \hat{\alpha}_{OLS}, \Sigma, y) \times W(\Sigma^{-1} | y, \hat{\alpha}_{OLS}, T - (mp + 1) - m - 1) \\ \alpha | \Sigma, y &\sim N(\hat{\alpha}, \Sigma \otimes (X'X)^{-1}) \\ \Sigma^{-1} | y &\sim W(S^{-1}, T - (mp + 1) - m - 1) \end{aligned}$$

where $\Sigma \otimes (X'X)^{-1}$ is the OLS estimator of $var(\hat{\alpha})$.

- $W(\cdot)$ is a Wishart distribution where $T - (mp + 1) - m - 1$ are degrees of freedom. S is the sum of squared errors computed using the OLS estimates of $\hat{\alpha}$.
- Define also the $T \times m$ matrix

$$Y = [(y_{11} \dots y_{1T})', \dots, (y_{m1} \dots y_{mT})'].$$

- Using $vec(\hat{A}) = \hat{\alpha}$, the OLS estimates are $\hat{A} = (X'X)^{-1}(X'Y)$ and $S = (Y - X\hat{A})'(Y - X\hat{A})$.



Bayesian estimation of Vector Autoregressions IV

- We know how to compute the likelihood function of a VAR, but we also need to build a prior for the parameters α and Σ such that we can then compute a posterior density. Based on the posterior density we can compute the conditional mean $E(\alpha | y)$ and the conditional variance $var(\alpha | y)$ that deliver the usual estimates and standard errors.



The Minnesota Prior I

- The assumption about prior densities (distributions) is important because the prior determines whether the posterior distribution can be computed analytically or numerically (using Markov-Chain Monte Carlo methods).
- The Minnesota prior delivers analytical solution for the posterior distribution. It is based on early work of Bayesian econometricians (specially Litterman) in Minnesota (Fed).
- The Minnesota prior is a **shrinkage prior** since its main role is to shrink coefficients toward zero.
- The prior about Σ is very simple: $\Sigma = \hat{\Sigma}_{(OLS)}$, that is, the OLS estimate of the variance-covariance matrix.



The Minnesota Prior II

- The prior distribution of interest is then:

$$\alpha \sim N(\underline{\alpha}, \underline{V}).$$

- Litterman assumption is that macroeconomic time series (log levels) resemble a random walk ($y_{1t} = y_{1t-1} + \varepsilon_{1t}$), so values in $\underline{\alpha}$ are all zero except when $i = j, \alpha_{ii} = 1$.
- However, this is not adequate when using a VAR of growth rates. A better suggestion is:

$$\underline{\alpha} = 0_{m(mp+1)},$$

which ensures shrinkage of the VAR coefficients toward zero, eliminating overfitting.



The Minnesota Prior III

- The prior for \underline{V} should be also based on the prior of Σ . Denote σ_{ij} the elements of the $m \times m$ matrix Σ . Call p_s the correspondent lag length of the specific coefficient, and that \underline{V}_i is related to the elements of \underline{V} of equation i .

$$\underline{V}_i = \begin{cases} \frac{a_1}{p_s^2} & \text{for coefficients on own lags} \\ \frac{a_2 \sigma_{ii}}{p_s^2 \sigma_{jj}} & \text{for coefficients on lags of variable } j \neq i \end{cases}$$

- The intuition is that as the lag length increases, coefficients are increasingly shrunk towards zero and, if $a_1 > a_2$, own lags are more likely to be important predictors than lags of other variables. The value of a_1 and a_2 depend on the empirical application and the values of σ_{ii} can be obtained from $\hat{\Sigma}_{(OLS)}$.



The Minnesota Prior IV

- The advantage of this prior is that the posterior distribution is also normal:

$$\alpha|y \sim N(\bar{\alpha}, \bar{V})$$

with conditional mean:

$$\bar{\alpha} = \bar{V} \left[\underline{V}^{-1} \underline{\alpha} + \left(\hat{\Sigma}^{-1} \otimes X \right)' y \right]$$

and conditional variance:

$$\bar{V} = \left[\underline{V}^{-1} + \left(\hat{\Sigma}^{-1} \otimes (X'X) \right) \right]^{-1}$$



The Minnesota Prior V

- Litterman (1986) shows that Bayesian VAR models estimated with the Minnesota prior deliver good forecasting performance. Forecasts are computed based on the conditional mean estimates of the parameters.
- Bayesian VARs can be also employed for impulse-reponse analysis. Identification strategies and issues are similar to the ones of the frequentist approach.



Example of Random Walk Minnesota Prior

