The Econometrics of DSGE Models

Giuseppe Ragusa
Luiss University

EIEF
Lecture 6: (Bayesian) VAR

March 16, 2017

Vector Autoregressions

- $VAR(p)$

$$y_t = \underset{n\times 1}{C} + \underset{n\times n}{A_1} y_{t-1} + \ldots + A_p y_{t-p} + \underset{n\times 1}{e_t}$$

- Flexible multivariate model
- Bridge between reduced-form and structural models

Vector Autoregressions

Notation (I)

Rewrite the VAR as

$$\underset{(1\times n)}{y_t} = \underset{(1\times q)}{z_t}\underset{(q\times n)}{\Gamma} + e_t$$

where $q = np + 1$ and

$$z_t = (1, y'_{t-1}, \ldots, y'_{t-p})$$
$$\Gamma = (C'\, A'_1 \ldots A'_p)$$

Stacking along the time dimension we can write

$$\underset{(T\times n)}{Y} = \underset{(T\times q)}{Z}\Gamma + \underset{(T\times n)}{E}$$

where

$$Y \equiv (y_1\, y_2 \ldots y_T) \quad E \equiv (e_1\, e_2 \ldots e_T)$$

Vector Autoregressions

Notation (II)

We can also write the VAR as

$$\begin{aligned} vec(Y) &= vec(Z\Gamma + E) \\ &= vec(Z\Gamma) + vec(E) \\ &= (I_n \otimes Z) vec(\Gamma) + vec(E) \end{aligned}$$

$$\boxed{\underset{(nT\times 1)}{y} = \underset{(nT\times qn)}{(I_n \otimes Z)} \underset{(qn\times 1)}{\beta} + e}$$

where

$$y = vec(Y), \beta \equiv \mathrm{vec}(\Gamma), u = vec(U).$$

Vector Autoregressions

Notation IV

Let

$$\xi_t = \begin{bmatrix} y_t - C \\ y_{t-1} - C \\ \vdots \\ \vdots \\ y_{t-p+1} - C \end{bmatrix}, \; F = \begin{bmatrix} A_1 & A_2 & A_3 & \cdots & A_{p-1} & A_p \\ I_n & 0 & 0 & \cdots & \cdots & 0 \\ 0 & I_n & 0 & \cdots & 0 & 0 \\ & & \ddots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & I_n & 0 \end{bmatrix}, \; v_t = \begin{bmatrix} e_t \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}.$$

Then, the VAR(p) can be written as VAR(1) in ξ_t

$$\xi_t = F\xi_{t-1} + v_t$$

Assumptions

- Variance

$$\begin{aligned} E[u_t u_t'] &= \Sigma \\ E[uu'] &= \Sigma \otimes I_T \\ E[uu'|z] &= \Sigma \otimes I_T \end{aligned}$$

- Distribution

$$u \sim N(0, \Sigma \otimes I_T)$$

or, using the matrix notation

$$U \sim MN(0, I_T, \Sigma)$$

where $MN(M, R, C)$ denote the **matric-variate** normal distribution with mean M, row-wise variance R and column wise variance C.

Covariance Stationarity - MA(∞)

From $\xi_t = F\xi_{t-1} + u_t$, we have

$$\xi_{t+s} = v_{t+s} + Fv_{t+s-1} + F^2 v_{t+s-2} + \ldots + F^{s-1} v_{t+1} + F^s \xi_t.$$

Definition

A VAR(p) is covariance stationary if the eigenvalues of the matrix F satisfies

$$|I_n \lambda^p - B_1 \lambda^{p-1} - B_2 \lambda^{p-2} - \cdots - B_p| = 0$$

In this case, $F^s \to 0$, as $s \to \infty$. Thus,

$$y_t = C + u_t + \Psi_1 u_{t-1} + \Psi_2 u_{t-2} + \cdots = C + \Psi(L) u_t$$

where

$$\Psi_j = F_{11}^{(j)}$$

and

$$\Psi(L) = I_n - \Psi_1 L - \Psi_2 L^2 - \ldots$$

Impulse-Response Function

For

$$y_t = C + u_t + \Psi_1 u_{t-1} + \Psi_2 u_{t-2} + \cdots$$

the matrix Ψ_s has the interpretation

$$\frac{\partial y_{t+s}}{\partial u_t'} = \Psi_s.$$

Thus, the row i, column j element of Ψ_s identifies the consequences of a one unit increase in the jth variable's innovation at date t (u_{jt}) for the value of the ith variable at time $t+s$ $(y_{i,t+s})$

$$\frac{\partial y_{i,t+s}}{\partial u_{jt}} = \Psi_{ij}.$$

Impulse-response Function

A plot of

$$\frac{\partial y_{i,t+s}}{\partial u_{jt}} = \Psi_{ij},$$

as a function of s is called the *impulse-response function.* It describes the response of $y_{j,t+s}$ to a one-time impulse in y_{jt} with all the other variables dated t or earlier held constant.

- Is there a sense in which this multiplier can be viewed as measuring the casual effect of y_j on y_i? (Not really!)

Classical inference

Since

$$F_{11}^{(j)} = f(C, A_1, \ldots, A_p),$$

we need to estimate $\Gamma = (C', A_1', \ldots, A_p')$

$$Y = Z\Gamma + E$$

- We can estimate $\beta = vec(\Gamma)$ by OLS

$$\hat{\Gamma} = (Z'Z)^{-1}Z'Y,$$
$$\hat{\beta} = \mathrm{vec}(\hat{\Gamma})$$
$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Omega),$$
$$\Omega = \Sigma \otimes E(z_t' z_t)^{-1}$$

- Since each equations has the same regressors

$$OLS = SUR = MLE$$

Bayesian inference

Bayesian inference will be based on the posterior distribution

$$p(\beta, \Sigma | y) \propto \underbrace{p(y|\beta, \Sigma)}_{\text{likelihood}} \underbrace{p(\beta|\Sigma)p(\Sigma)}_{\text{prior}}$$

- What is the likelihood?
- What are appropriate priors?

Minnesota prior

Consider a reduced form VAR(p):

$$Y_t = c + A_1 Y_{t-1} + A_2 Y_{t-2} + \ldots + A_p Y_{t-p} + e_t \text{ where } e_t \sim WN(\underline{0}, \Sigma)$$

- Litterman(1979, 1980): precision of estimates and forecasting performance can be improved by incorporating 'restrictions' in the form of a prior distribution on the parameters.
- Prior based on stylized facts about his data: macroeconomic variables for the US that could be well characterized by unit root processes
  ⇒ he proposed shrinking towards a univariate random walk for each variable in the VAR (RW: $y_t = y_{t-1} + e_t$)
  ⇒ prior on $(A_k)_{ij}$: coefficient in position (i, j) of matrix A_k.
- The error variances are assumed to be known, $Var(e_{t,j}) = \sigma_j^2$, where σ_j^2 is the OLS residual variance for equation j in the VAR.

I. The shrinkage towards univariate random walks corresponds the setting the prior mean of A to:

$$\gamma_{ij} = \mathbb{E}[(A_k)_{ij}] = \begin{cases} 1 & \text{if } i = j \text{ and } k = 1 \\ 0 & \text{otherwise} \end{cases}$$

Shrinkage towards zero for longer lags, $k > 1$, reflecting the prior notion that more distant observations are less influential.

II. In addition, a different amount of shrinkage is applied to lags of the dependent variable than to lags of other variables in the same equation. Typically more shrinkage is applied to lags of other variables to reinforce the univariate random walk nature of the prior. This is done setting the prior standard deviations to:

$$\tau_{ij} = \mathbb{V}[(A_k)_{ij}] = \begin{cases} \frac{\lambda^2}{k^2} & j = i \\ \frac{\lambda^2}{k^2}\frac{\sigma_i^2}{\sigma_j^2} & \text{otherwise} \\ \infty & \text{deterministic variable} \end{cases} .$$

λ controls the overall tightness, typical value used 0.2

Thus the prior: $p(A_{ij}) \sim N(\gamma_{ij}, \tau_{ij})$

Variations on Minnesota prior

- **Stationary variables**: For variables believed to be stationary the prior mean on $k = 1$ can be set to a value less than 1.
- **'Exogenous' variables**: for $i \neq j$ $\tau_{ij} = (\sigma_i^2 \lambda_1 \lambda_2 / \sigma_j^2 k^{\lambda_3})$, $\lambda_1 \lambda_2$ is more flexible w.r.t. use only λ^2, λ_3 before fix to 2.
- **Deterministic term**: for the constant use a propter prior $\tau_{ij} = \sigma_i^2 \lambda_4$ while still being uninformative about by setting λ_4 (moderately) large.
- **Sum of coefficients prior**: expresses the prior notion that the sum of coefficients on own lags is 1 and the sum of coefficients on the lags of each of the other variables is 0 as well as the idea that the recent average of the variable should be a reasonable forecast.
- **Dummy initial observations prior**: This prior also implies that the initial observations is a good forecast without enforcing specific parameter values and induces prior correlation among all parameters in the equation.

Extensions...

Now we will study 3 extensions of standard Minnesota prior:

- Normal-Wishart
- Independent Normal-Wishart
- Hierarchical

For notational simplicity, we consider the compact form:

$$Y = Z\Gamma + E$$

where

- $\Gamma = (c, A_1', , \ldots, A_p')$
- $e_t \sim N(0, \Sigma)$
- $k = np + 1$

Conjugate prior: Normal-Wishart

Kadiyala and Karlsson (1993,1997) relax the assumption of a known Σ, and studies the effect of varying the family of distribution used to parametrize the prior beliefs.
We focus on conjugate normal-Wishart prior:

- Normal-Wishart prior is the natural conjugate prior for normal multivariate regressions.
- It generalizes the original Litterman prior by treating the error variance-covariance matrix, Σ, as an unknown positive definite symmetric matrix rather than a fixed diagonal matrix.

Adding and subtracting $Z\hat{\Gamma}$ in the likelihood ($\hat{\Gamma} = (Z'Z)^{-1}Z'Y$) and letting $S = (Y - Z\hat{\Gamma})(Y - Z\hat{\Gamma})$ be the error sum of squares:

$$
\begin{aligned}
L(Y|\Gamma,\Sigma) \propto |\Sigma|^{-T/2} \exp\left\{ -\frac{1}{2} tr\left(\Sigma^{-1} S\right) \right\} \\
\times \exp\left\{ -\frac{1}{2} tr\left[\Sigma^{-1} \left(\Gamma - \hat{\Gamma}\right)' Z'Z \left(\Gamma - \hat{\Gamma}\right) \right] \right\}
\end{aligned}
$$

has the form of a normal-Wishart distribution (as a function of Γ and Σ).

Specifying the prior

$$\Gamma|\Sigma \sim MN_{kn}(\underline{\Gamma}, \Sigma, \underline{\Omega})$$

$$\Sigma \sim iW(\underline{S}, \underline{\nu})$$

The corresponding marginal posterior:

$$\text{(P1)} \quad \Gamma|Y_T, \Sigma \sim MN_{kn}(\overline{\Gamma}, \Sigma, \overline{\Omega})$$

where

$$\overline{\Omega}^{-1} = \underline{\Omega}^{-1} + ZZ'$$

$$\overline{\Gamma} = \overline{\Omega}\left(\underline{\Omega}^{-1}\underline{\Gamma} + Z'Z\hat{\Gamma}\right)$$

and

$$\text{(P2)} \quad \Sigma|Y_T \sim iW(\overline{S}, \overline{\nu})$$

where

$$\overline{\nu} = T + \underline{\nu}$$

$$\overline{S} = \underline{S} + S + (\underline{\Gamma} - \hat{\Gamma})'(\underline{\Omega} + (Z'Z)^{-1})^{-1}(\underline{\Gamma} - \hat{\Gamma})$$

Simulating from the posterior distribution: With a normal-Wishart posterior we can proceed as in Algorithm I using the posterior distributions (P1) and (P2).

Ancillary note: matrix-variate normal

The $p \times q$ matrix X is said to have a matrix-variate normal distribution

$$\underset{p\times q}{X} \sim MN_{pq}(M, Q, P)$$

where M is $(p \times q)$, and $P_{(p\times p)}, Q_{(q\times q)}$ are positive semidefinite matrix, if $vec(X)$ is multivariate normal

$$vec(X) \sim N(vec(M), Q \otimes P)$$

Algorithm I. Simulating the predictive distribution with a normal-Wishart posterior

For $j = 1, \dots, R$

1. Generate $\Sigma^{(j)}$ from the marginal posterior $\Sigma | Y_T \sim iW(\overline{S}, \overline{v})$
2. Generate $\Gamma^{(j)}$ from the conditional posterior $\Gamma | Y_T, \Sigma^{(j)} \sim N(\overline{\Gamma}, \Sigma, \overline{\Omega})$
3. Generate $u^{(j)}_{T+1}, \dots, u^{(j)}_{T+H}$ from $u_t \sim N(0, \Sigma^{(j)})$ and calculate recursively:

$$\widetilde{Y}^{(j)}_{T+h} = c^{(j)} + \sum_{i=1}^{h-1} \widetilde{Y}_{T+h-1} A_i^{(j)} + \sum_{i=h}^{p} Y_{T+h-1} A_i^{(j)} + u^{(j)}_{T+h}$$

Independent Normal-Wishart

In the independent Normal-Wishart the joint prior is:

$$p(\Gamma, \Sigma) = p(\Gamma)p(\Sigma)$$

whereas in the Normal-Wishart is: $p(\Gamma, \Sigma) = p(\Gamma|\Sigma)p(\Sigma)$.
As prior we assume:

$$\gamma \sim N(\underline{\gamma}, \underline{\Sigma}_{\gamma})$$
$$\Sigma \sim iW(\underline{S}, \underline{\nu})$$

where $\gamma = vec(\Gamma)$.

The corresponding posterior:

$$\text{(P3)} \quad \gamma | Y_T, \Sigma \sim N(\overline{\gamma}, \overline{\Sigma}_\gamma)$$

where

$$\overline{\Sigma}_\gamma = (\underline{\Sigma}_\gamma^{-1} + \Sigma^{-1} \otimes ZZ')^{-1}$$

$$\overline{\gamma} = \overline{\Sigma} \left(\underline{\Sigma}^{-1} \underline{\gamma} + (\Sigma^{-1} \otimes Z'Z) \hat{\gamma} \right)$$

and

$$\text{(P4)} \quad \Sigma | Y_T \sim iW(\overline{S}, \overline{v})$$

where

$$\overline{v} = T + \underline{v}$$

$$\overline{S} = \underline{S} + (Y - Z\hat{\Gamma})'(Y - Z\hat{\Gamma})$$

Simulating from the posterior distribution: With the full conditional posteriors in hand a straightforward Gibbs sampling scheme is available for sampling from the posterior and predictive distributions (see Algorithm II). Kadiyala and Karlsson (1997) show that the Gibbs sampler convergences quickly to the posterior distribution and a few hundred draws may be sufficient as burn-in when the posterior is unimodal.

Algorithm II. Gibbs sampler for independent normal-Wishart priors

Select a starting value, $\gamma^{(0)}$ for γ. For $j = 1, \ldots, B + R$:

1. Generate $\Sigma^{(j)}$ from the full conditional posterior (P4) with $\overline{S}$ evaluated at $\gamma^{(j-1)}$.
2. Generate $\gamma^{(j)}$ from the full conditional posterior (P3) with $\overline{\Sigma}_\gamma$ evaluated at $\Sigma^{(j))}$.
3. If $j > B$, generate $u^{(j)}_{T+1}, \ldots, u^{(j)}_{T+H}$ from $u_t \sim N(0, \Sigma^{(j)})$ and calculate recursively:

$$\widetilde{Y}^{(j)}_{T+h} = c^{(j)} + \sum_{i=1}^{h-1} \widetilde{Y}_{T+h-1} A^{(j)}_i + \sum_{i=h}^{p} Y_{T+h-1} A^{(j)}_i + u^{(j)}_{T+h}$$

Hierarchical prior

- The parameters in Minnesota prior (a.k.a. hyperparameters) are generally chosen as:
  - default values similar to the ones used by Litterman or see Canova(2007)
  - to minimize the forecast errors over a training sample
- As an alternative Giannone et al. (2012) suggests a more flexible approach where one more layer is added to the prior structure by placing a prior on the hyperparameters in a hierarchical fashion:
  - MP classic: $\lambda \rightarrow p(B)$
  - MP hierarchical: $p(\lambda) \rightarrow \lambda \rightarrow p(B) \Rightarrow p(B) = \int p(B|\lambda)p(\lambda)d\lambda$

- Collecting the hyperparameters in the vector δ and working with the normal-Wishart family of prior distributions, the prior structure becomes:

  $$p(\gamma|\Sigma,\delta)p(\Sigma|\delta)p(\delta)$$

  where
  - $\Sigma \sim iW(\underline{S},\underline{v})$.
  - $\gamma|\Sigma \sim N(\underline{\gamma},\Sigma\otimes\Omega)$.
  - the elements $\underline{S},\underline{v},\underline{\gamma},\Omega$ are function of δ.

- **Basic set up:** Original Minnesota Prior $\Rightarrow$ the hyperprior is λ. Where $p(\lambda)\sim\Gamma$ with mode 0.2.
- **Extended** set up in the paper: MP with extension + dummy variable +sum of coefficients

Giannone et al. 2012 conclusion:

- ‘intermediate’ prior (neither flat nor dogmatic) produce better forecast performance
- For large database is not possible use flat priors
- $\uparrow$ database dimension $\Rightarrow \uparrow$ tightness

Simulating from the posterior distribution: The joint posterior of Γ, Σ and δ is not available in closed form but Giannone et al. (2012) devise a Metropolis-Hastings sampler for the joint distribution (see Algorithm III). The algorithm generates δ from the marginal posterior with a Metropolis-Hastings update, after convergence of the δ sampler Σ and Γ can be drawn from their distributions conditional on δ.

Algorithm III MCMC sampler for a VAR with hierarchical prior

Select starting values for the hyperparameters $\delta^{(0)}$ (n.b. Giannone et al. (2012) suggests using the posterior mode of δ). For $j = 1, \ldots, B + R$

1. Sample $\delta^{(j)}$:
   - 1.a Draw a proposal, $\delta^{\star}$, for the hyperparameters from the random walk proposal distribution, $\delta^{\star} \sim N(\delta^{(j-1)}, cH^{-1})$ where H is the Hessian of the negative of the logposterior for δ and c is set to achieve approximately 20% acceptance rate.
   - 1.b Compute

     $$\alpha^{(j)} = min\left\{1, \frac{p(Y|\delta^{\star})p(\delta^{\star})}{p(Y|\delta^{(j-1)})p(\delta^{(j-1)})}\right\}$$

     where $p(Y|\delta^{\star})$ is the marginal likelihood of Y (which is a matrix-variate T-Student).
   - 1.c Sample $u \sim U[0,1]$
   - 1.d If $\alpha^{(j)} > u$ then $\delta^{(j)} = \delta^{\star}$, otherwise set $\delta^{(j)} = \delta^{(j-1)}$

$\Rightarrow$ Redo step 1 if $j < B$, otherwise continue

2. Draw $\Sigma^{(j)}$ from the full conditional posterior (P2) $\Sigma | Y_T, \delta^{(j)}$
3. Draw $\Gamma^{(j)}$ from the full conditional posterior (P1) $\Gamma | Y_T, \Sigma^{(j)}, \delta^{(j)}$
4. If $j > B$, generate $u^{(j)}_{T+1}, \ldots, u^{(j)}_{T+H}$ from $u_t \sim N(0, \Sigma^{(j)})$ and calculate recursively:

$$\widetilde{Y}^{(j)}_{T+h} = c^{(j)} + \sum_{i=1}^{h-1} \widetilde{Y}_{T+h-1} A_i^{(j)} + \sum_{i=h}^{p} Y_{T+h-1} A_i^{(j)} + u^{(j)}_{T+h}$$

Large BVAR

- Because of the problem with having to estimate so many parameters, most VAR papers have tended to use a small number of macroeconomic variables.
- However, economists in engaged in forecasting tend to look at a huge range of variables that are available at a monthly or weekly frequency.
- Based on the theoretical results of De Mol et al. (2008), Banbura et al. (2010) suggest that the degree of shrinkage applied through the prior should increase with the size of the model.
- Working with a normal-Wishart prior distribution with Minnesota type prior beliefs, the overall scaling factor λ (slide 13) determines the amount of shrinkage.

- Model $\mathscr{M}$: small (3), CEE (7), medium (20), large (131).
- For a desired fit, λ is chosen as

$$\lambda_{\mathscr{M}}(\mathrm{fit}) = \arg\min_{\lambda} |\mathrm{fit} - \frac{1}{3}\sum_{i\in I}\frac{\mathrm{msfe}_i^{\lambda,\mathscr{M}}}{\mathrm{msfe}_i^0}|$$

  where msfe is the in-sample one-step ahead mean square forecast error evaluated over a training sample and $I = \{\text{Employment, CPI, FedFunds}\}$
- In the main paper they report the results where the desired fit coincides with the one obtained by OLS estimation ($\lambda = \infty$) on the small model with $p = 13$:

$$|\frac{1}{3}\sum_{i\in I}\frac{\mathrm{msfe}_i^{\lambda,\mathscr{M}}}{\mathrm{msfe}_i^0}|_{\lambda=\infty,\mathscr{M}=\mathrm{small}}$$

Table 1: Relative MSFE, BVAR

		SMALL	CEE	MEDIUM	LARGE
h=1	EMPL	1.14	0.67	0.54	0.46
	CPI	0.89	0.52	0.50	0.50
	FFR	1.86	0.89	0.78	0.75
h=3	EMPL	0.95	0.65	0.51	0.38
	CPI	0.66	0.41	0.41	0.40
	FFR	1.77	1.07	0.95	0.94
h=6	EMPL	1.11	0.78	0.66	0.50
	CPI	0.64	0.41	0.40	0.40
	FFR	2.08	1.30	1.30	1.29
h=12	EMPL	1.02	1.21	0.86	0.78
	CPI	0.83	0.57	0.47	0.44
	FFR	2.59	1.71	1.48	1.93
λ		∞	0.262	0.108	0.035

Notes: Table reports MSFE relative to that from the benchmark model (random walk with drift) for employment (EMPL), CPI and federal funds rate (FFR) for different forecast horizons h and different models. SMALL, CEE, MEDIUM and LARGE refer to the VARs with 3, 7, 20 and 131 variables, respectively. λ is the shrinkage hyperparameter and is set so that the average in-sample fit for the three variable of interest is the same as in the SMALL model estimated by OLS. The evaluation period is 1971-2003.