Bayesian Vector Autoregressions

Giovanni Ricco

10th November 2023

Bayesian Vector Autoregressions

- **Univariate AR(p)** model – y_t is 1×1:

$$y_t = c + a_1 y_{t-1} + \cdots + a_p y_{t-p} + u_t \qquad u_t \sim \mathcal{N}(0, \sigma)$$

y_t is function of its lagged realisations and a stochastic innovation

- **VAR(p) model** – y_t is $n \times 1$:

$$y_t = C + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t \qquad u_t \sim \mathcal{N}(0, \Sigma)$$

where the A_j $(j = 1, \ldots, p)$ and Σ are $n \times n$ matrices, and C is $n \times 1$

Bayesian Vector Autoregressions

- Let's write the **VAR(p) likelihood function**, **conditional** on the first p observations
- Re-write the VAR(p) as

$$y_t = \underbrace{[A_1 \dots A_p C]}_{A'} \underbrace{\begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-p} \\ 1 \end{pmatrix}}_{x_t} + u_t \qquad u_t \sim \mathcal{N}(0, \Sigma)$$

  that is

$$y_t = A' x_t + u_t$$

- Just a **multivariate regression model**!

Bayesian Vector Autoregressions

- The **conditional density** of y_t is

$$p(y_t|y_{t-1},\ldots,y_{t-p},A,\Sigma)$$
$$\propto |\Sigma|^{-1/2} exp\left\{-\frac{1}{2}(y_t - A'x_t)'\Sigma^{-1}(y_t - A'x_t)\right\}$$

- Note that: ① for a a vector $n \times 1$ and B a matrix $n \times n$

$$a'Ba = tr[aBa]$$

② Trace is invariant under **cyclic permutations**

$$tr\left[a'Ba\right] = tr\left[Baa'\right] = tr\left[aa'B\right]$$

- Hence

$$p(y_t|y_{t-1},\ldots,y_{t-p},A,\Sigma) \propto |\Sigma|^{-1/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(y_t - A'x_t)(y_t - A'x_t)'\right]\right\}$$

Bayesian Vector Autoregressions

- The **joint density** for the observations

$$Y_{1:T} \equiv [y_1, \dots, y_T]$$

conditional on the first p observations

$$Y_{-p+1:0} \equiv [y_{-p+1}, \dots, y_0]$$

is the **product of conditional densities**

$$\begin{aligned} p(Y_{1:T}|Y_{-p+1:0}, A, \Sigma) &= \prod_{t=1}^{T} p(y_t|Y_{-p+1:t-1}, A, \Sigma) \\ &= \prod_{t=1}^{T} p(y_t|Y_{t-p:t-1}, A, \Sigma) \\ &\propto \prod_{t=1}^{T} \left(|\Sigma|^{-1/2} exp\left\{ -\frac{1}{2} tr\left[\Sigma^{-1}(y_t - A'x_t)(y_t - A'x_t)' \right] \right\} \right) \end{aligned}$$

Bayesian Vector Autoregressions

- Since $tr[A] + tr[B] = tr[A + B]$, we can write

$$
\begin{aligned}
& p(Y_{1:T}|Y_{-p+1:0}, A, \Sigma) \\
& \qquad \propto \prod_{t=1}^{T} \left(|\Sigma|^{-1/2} exp \left\{ -\frac{1}{2} tr \left[\Sigma^{-1}(y_t - A'x_t)(y_t - A'x_t)' \right] \right\} \right) \\
& \qquad \propto |\Sigma|^{-T/2} exp \left\{ -\frac{1}{2} tr \left[\Sigma^{-1} \sum_{t=1}^{T} (y_t - A'x_t)(y_t - A'x_t)' \right] \right\}
\end{aligned}
$$

Bayesian Vector Autoregressions

- Now define

$$Y = \begin{pmatrix} y_1' \\ \vdots \\ y_T' \end{pmatrix} \qquad X = \begin{pmatrix} x_1' \\ \vdots \\ x_T' \end{pmatrix}$$

$$p(Y_{1:T} | Y_{-p+1:0}, A, \Sigma) \propto |\Sigma|^{-T/2} exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}(Y - XA)'(Y - XA)\right]\right\}$$

Bayesian Vector Autoregressions

- As done before, define the OLS estimator

$$\widehat{A} = (X'X)^{-1}X'Y$$

and the sum of squared OLS residual matrix

$$\widehat{S} = (Y - X\widehat{A})'(Y - X\widehat{A})$$

- as in the univariate regression

$$(Y - XA)'(Y - XA) = \widehat{S} + (A - \widehat{A})'X'X(A - \widehat{A})$$

- hence

$$p(Y_{1:T}|Y_{-p+1:0}, A, \Sigma) \propto |\Sigma|^{-T/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\widehat{S}\right]\right\}$$
$$\times exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(A - \widehat{A})'X'X(A - \widehat{A})\right]\right\}$$

Bayesian Vector Autoregressions

- Using the following matrix results

$$(A \otimes B)' = (A' \otimes B')$$

$$(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$$

$$tr[A'BCD'] = vec(A)'(D \otimes B)vec(C)$$

- we get

$$p(Y_{1:T}|Y_{-p+1:0}, A, \Sigma) \propto |\Sigma|^{-T/2} exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\widehat{S}\right]\right\}$$
$$\times exp\left\{-\frac{1}{2} vec(A - \widehat{A})'[\Sigma \otimes (X'X)^{-1}]^{-1} vec(A - \widehat{A})\right\}$$

Non-informative Priors

- The **posterior distribution**

$$p(A,\Sigma|Y) = p(A|\Sigma,Y)p(\Sigma|Y) = \underbrace{p(Y_{1:T}|Y_{-p+1:0},A,\Sigma)}_{likelihood}\underbrace{p(A|\Sigma)p(\Sigma)}_{prior}$$

- With **non-informative priors** on A and Σ

$$p(vec(A)|\Sigma) \propto 1$$

$$p(\Sigma) \propto |\Sigma|^{-\frac{n+1}{2}}$$

- The posterior conditional distributions are

$$vec(A)|Y,\Sigma \sim \mathcal{N}\left(vec(\widehat{A}),\Sigma \otimes (X'X)^{-1}\right)$$

$$\Sigma|Y \sim \mathcal{IW}\left(\widehat{S},T-k\right)$$

Matricvariate Normal Distribution, and Inverse Wishart

Informative priors

- A general **Normal-Inverted Wishart** prior has the form:

$$vec(A)|\Sigma \sim \mathcal{N}(vec(A_0), \Sigma \otimes \Omega_0)$$

$$\Sigma \sim \mathcal{IW}(S_0, \nu_0)$$

conjugate priors!

- How to set prior parameters A_0, Ω_0, S_0 and ν_0?

- Most used macro-priors: Minnesota priors (see Doan, Litterman and Sims, 1994)

Informative priors

What do we know a priori about macro variables?

Minnesota priors

- **Prior model:** each variable i is an independent **random walk** process

$$y_{i,t} = c + y_{i,t-1} + u_{i,t}$$

- ... or more generally a first order independent autoregressive process

$$y_{i,t} = c + \delta_i y_{i,t-1} + u_{i,t}$$

Minnesota priors

- These prior beliefs are imposed by setting the following moments for the prior distribution of the coefficients (conditional on Σ)

$$\mathbb{E}[(A_k)_{ij}|\Sigma] = \begin{cases} \delta_i & j = i, k = 1 \\ 0 & otherwise \end{cases} \tag{1}$$

$$\mathbb{V}[(A_k)_{ij}|\Sigma] = \frac{\lambda_1^2}{k^2}\frac{\Sigma_{ij}}{\sigma_j^2} \tag{2}$$

- λ_1 is the parameter setting overall tightness of the priors

Minnesota priors

- These prior beliefs are imposed by setting the following moments for the prior distribution of the coefficients (conditional on Σ)

$$\mathbb{E}[(A_k)_{ij}|\Sigma] = \begin{cases} \delta_i & j = i, k = 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\mathbb{V}[(A_k)_{ij}|\Sigma] = \frac{\lambda_1^2}{k^{2\lambda_2}} \frac{\Sigma_{ij}}{\sigma_j^2} \tag{2}$$

- λ_1 is the parameter setting overall tightness of the priors
- λ_2 sets the increase of tightness at longer lags

Minnesota priors

- The coefficients

$$A_1, ..., A_p$$

are assumed to be a priori **independent and normally distributed**

- The parameters prior on the covariance matrix of the residuals, S_0 and ν_0 are chosen by imposing that

$$\mathbb{E}[\Sigma] = \frac{S_0}{\nu_0 - n - 1}$$

exists and matches a given diagonal covariance matrix

$$\frac{S_0}{\nu_0 - n - 1} = diag(\sigma_1^2, \ldots, \sigma_n^2)$$

- The prior on the intercept is diffuse

Dummies to Implement the Minnesota Priors

$$y_d = \begin{pmatrix} diag(\delta_1\sigma_1, \dots, \delta_n\sigma_n)/\lambda_1 \\ 0_{n(p-1)\times n} \\ \dots\dots\dots\dots \\ diag(\sigma_1, \dots, \sigma_n) \\ \dots\dots\dots\dots \\ 0_{1\times n} \end{pmatrix}$$

$$x_d = \begin{pmatrix} J_p \otimes diag(\sigma_1, \dots, \sigma_n)/\lambda_1 & 0_{np\times 1} \\ \dots\dots\dots\dots & \dots \\ 0_{n\times np} & 0_{n\times 1} \\ \dots\dots\dots\dots & \dots \\ 0_{1\times np} & \epsilon \end{pmatrix}$$

where $J_p = diag(1, 2, \dots, p)$

Dummies to Implement the Minnesota Priors

$$y_d = \begin{pmatrix} diag(\delta_1\sigma_1, \ldots, \delta_n\sigma_n)/\lambda_1 \\ 0_{n(p-1)\times n} \\ \ldots\ldots\ldots\ldots \\ diag(\sigma_1, \ldots, \sigma_n) \\ \ldots\ldots\ldots\ldots \\ 0_{1\times n} \end{pmatrix}$$

$$x_d = \begin{pmatrix} J_p \otimes diag(\sigma_1, \ldots, \sigma_n)/\lambda_1 & 0_{np\times 1} \\ \ldots\ldots\ldots\ldots\ldots & \\ 0_{n\times np} & 0_{n\times 1} \\ \ldots\ldots\ldots\ldots\ldots & \\ 0_{1\times np} & \epsilon \end{pmatrix}$$

More general form: $J_p = diag(1^{\lambda_2}, 2^{\lambda_2}, \ldots, p^{\lambda_2})$

Dummies to Implement the Minnesota Priors

- The **first block** of dummies imposes prior beliefs on the **autoregressive coefficients**
- The **second block** implements the prior for the **covariance matrix**
- The **third block** reflects **a very diffuse prior for the intercept** to be around zero

$$\epsilon \approx 0$$

Dummies to Implement the Minnesota Priors

Remark:

- Parameters should be set using **only prior knowledge**!
- However, it is common practice to set the scale parameters σ_i^2 using sample information
- For example, the variance of the **residuals of univariate autoregressive models** of order p for each variables y_{it}
- It is possible to do better...

Dummies to Implement the Minnesota Priors

$$y_d = x_d A + u_d$$

Example ($n = 2$, $p = 2$):

- The first n dummies impose priors on A_1

$$\begin{pmatrix} \frac{\delta_1 \sigma_1}{\lambda_1} & 0 \\ 0 & \frac{\delta_2 \sigma_2}{\lambda_1} \end{pmatrix} = \begin{pmatrix} \frac{\sigma_1}{\lambda_1} & 0 & 0 & 0 & 0 \\ 0 & \frac{\sigma_2}{\lambda_1} & 0 & 0 & 0 \end{pmatrix} A + \begin{pmatrix} u^d_{1,1} & u^d_{2,1} \\ u^d_{1,2} & u^d_{2,2} \end{pmatrix}$$

Dummies to Implement the Minnesota Priors

- The first observation implies:

$$
\begin{aligned}
\frac{\delta_1 \sigma_1}{\lambda_1} = \frac{\sigma_1}{\lambda_1} A_{1,11} + u^d_{1,1} \quad &\Longrightarrow \quad A_{1,11} = \delta_1 - \frac{u^d_{1,1}\lambda_1}{\sigma_1} \\
&\Longrightarrow \quad A_{1,11} \sim \mathcal{N}\left(\delta_1, \frac{\Sigma_{1,1}\lambda_1^2}{\sigma_1^2}\right) \\
0 = \frac{\sigma_1}{\lambda_1} A_{1,21} + u^d_{2,1} \quad &\Longrightarrow \quad A_{1,21} = -\frac{u^d_{2,1}\lambda_1}{\sigma_1} \\
&\Longrightarrow \quad A_{1,21} \sim \mathcal{N}\left(0, \frac{\Sigma_{2,1}\lambda_1^2}{\sigma_1^2}\right)
\end{aligned}
$$

- Prior tightness depends on the hyperparameter λ_1
- The smaller λ_1, the smaller the prior variance

Dummies to Implement the Minnesota Priors

- Dummies for the other lag ($p = 2$)

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 2^{\lambda_2}\sigma_1/\lambda_1 & 0 & 0 \\ 0 & 0 & 0 & 2^{\lambda_2}\sigma_2/\lambda_1 & 0 \end{pmatrix} A + \begin{pmatrix} u^d_{1,1} & u^d_{2,1} \\ u^d_{1,2} & u^d_{2,2} \end{pmatrix}$$

$$0 = (2^{\lambda_2}\sigma_1/\lambda_1)A_{2,11} + u^d_{1,1} \implies A_{2,11} = -\frac{u^d_{1,1}\lambda_1}{2^{\lambda_2}\sigma_1}$$

$$\implies A_{2,11} \sim \mathcal{N}\left(0, \frac{\Sigma_{1,1}\lambda_1^2}{2^{2\lambda_2}\sigma_1^2}\right)$$

- Prior tightness **increases** with λ_2 (in addition to λ_1)
- ... and, for given λ_2, **with the lag order l**

Dummies to Implement the Minnesota Priors

- Prior dummies for the covariance matrix are implemented by (λ_3 replications of)

$$\begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} A + \begin{pmatrix} u^d_{1,1} & u^d_{2,1} \\ u^d_{1,2} & u^d_{2,2} \end{pmatrix}$$

- Note that λ_3 determines the weight for the prior on Σ
- Suppose that

$$Z_i \sim \mathcal{N}(0, \sigma^2)$$

An estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{\lambda_3} \sum_{i=1}^{\lambda_3} Z_i^2$$

The larger λ_3 , the more informative the estimator (the tighter the prior)

Dummies to Implement the Minnesota Priors

- Prior dummies for the intercept...
- Check yourself! (**Exercise**)

The Posterior

Regression model **augmented with the dummies**:

$$\underset{T_* \times n}{y_*} = \underset{T_* \times k}{x_*} \underset{k \times n}{A} + \underset{T_* \times n}{U_*},$$

where

$$T_* = T + T_d$$

$$y_* = (y', y_d')'$$

$$x_* = (x', x_d')$$

and

$$U_* = (u', u_d')'$$

The Posterior

The posterior has the form:

$$vec(A)|\Sigma, y \sim N\left(vec(\widetilde{A}), \Sigma \otimes (x_*'x_*)^{-1}\right)$$

$$\Sigma|y \sim \mathcal{IW}\left(\widetilde{\Sigma}, \nu\right)$$

with

$$\widetilde{A} = (x_*'x_*)^{-1}x_*'y_*$$

$$\widetilde{\Sigma} = (y_* - x_*\widetilde{A})'(y_* - x_*\widetilde{A})$$

$$\nu = T_d + T - k$$

Remark: The posterior mean of the coefficients is the OLS estimate for the regression of y_* on x_*

Large BVARs

Large BVARs

- BVARs can accommodate $N \sim 100$ variables!
- **Very first 'larger' VAR:** Leeper, Sims and Zha (1996)
- **Reference (Theory):** De Mol, Giannone, Reichlin ('Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?', JE 2008)
- **Reference (Application):** Bańbura, Giannone, and Reichlin (Large Bayesian VARs, JAE 2010) and Koop ('Forecasting with Medium and Large Bayesian VARs', JAE 2011) and Bańbura, Giannone, Lenza ('Conditional forecasts and scenario analysis with vector autoregressions for large cross-sections', IJF 2015)

Large BVARs

- Why do large BVARs work in terms of forecasting (and structural identification)?
- Doesn't the number of coefficients blow up, causing overfitting and erratic forecasts? ('curse of dimensionality')

Bayesian Regression and Principal Components

- Let us write our model as

$$y = x\beta + \epsilon$$

- Consider a a shrinkage prior on β

$$\beta_i \sim \mathcal{N}\left(0, \frac{\sigma_i^2}{\lambda^2} I_k\right)$$

Bayesian Regression and Principal Components

- We can always apply a linear transformation H (invertible) to the regressors

$$y = (xH)(H^{-1}\beta) + \epsilon = F\gamma + \epsilon$$

where

$$F = xH$$

$$\gamma = H^{-1}\beta$$

- We have

$$\gamma_i \sim \mathcal{N}\left(0, \frac{\sigma_i^2}{\lambda^2} H^{-1}(H^{-1})'\right)$$

Bayesian Regression and Principal Components

- Compute the variance matrix of the (demeaned) regressors and take its eigen-(value/vector) decomposition

$$\frac{x'x}{T} = VDV'$$

where D is diagonal and V is orthonormal ($VV' = V'V = I$)

- Consider $H = VD^{-1/2}$, and hence $H^{-1} = D^{1/2}V'$

Bayesian Regression and Principal Components

- Now $F = xH = xVD^{-\frac{1}{2}}$ are the (standardised) principal components of x

$$\frac{F'F}{T} = I$$

- Since $H^{-1}(H^{-1})'$

$$\gamma_i \sim \mathcal{N}\left(0, \frac{\sigma_i^2}{\lambda^2} D\right)$$

Bayesian Regression and Principal Components

- With flat priors nothing would change!
- With the shrinkage prior, the more important is the principal component (higher d_r) the less you shrink!

$$\gamma_{ir} \sim \mathcal{N}\left(0, \frac{\sigma_i^2}{\lambda^2} d_r\right)$$

- Symmetric shrinkage on the variables implies asymmetric shrinkage on the principal components where we shrink more the less relevant the principal component!

Bayesian Regression and Factors

- If $\lambda \propto T$ and X has a factor structure with R factors, then asymptotically (for $T \to \infty$)
$$d_r \propto T \qquad \text{for } r \leq R$$
while d_r is bounded for $r > R$

- All F_r other than the first R are killed by the shrinkage prior

- Bayesian regression (Large VARs) tends to capture the factors that explain most of the variation in the predictors

- Suitable for large number of predictors if there is substantial comovement among predictors

Large BVARs

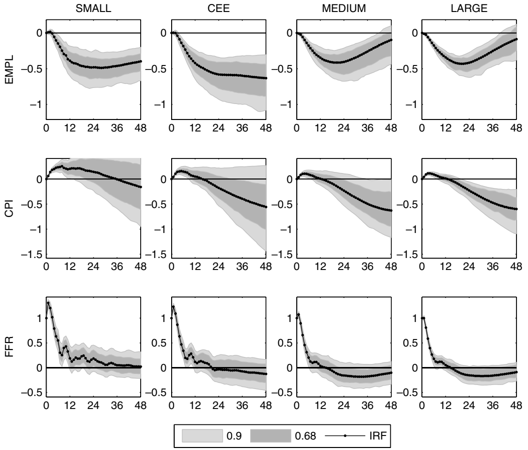

Figure: Monetary policy shock and the posterior coverage intervals at 0.68 and 0.9 level for employment (EMPL), CPI and federal funds rate (FFR). SMALL, CEE, MEDIUM and LARGE refer to VARs with 3, 7, 20 and 131 variables, respectively. (Banbura et al, 2010)