Bayesian VARs

Andrea Carriero

January 2018

The Bivariate VAR Model

- Suppose we have observed two time series, $Y_{1,t}$ and $Y_{2,t}$, over time.
- A priori, we wish to allow for the two time series to co-move. That is, past (lagged) values of $Y_{2,t}$ may potentially explain the current value of $Y_{1,t}$, and vice versa.
- **Bivariate VAR**(p)**:** A natural tool to model the *joint dynamics* of $(Y_{1,t}, Y_{2,t})$ is by extending the idea of ADL models as follows:

$$\begin{aligned} Y_{1,t} &= \mu_1 + \sum_{k=1}^{p} \phi_{1,k} Y_{1,t-k} + \sum_{k=1}^{p} \gamma_{1,k} Y_{2,t-k} + \varepsilon_{1,t}, \\ Y_{2,t} &= \mu_2 + \sum_{k=1}^{p} \phi_{2,k} Y_{2,t-k} + \sum_{k=1}^{p} \gamma_{2,k} Y_{1,t-k} + \varepsilon_{2,t}. \end{aligned}$$

- Separately, each of the two equations constitutes a restricted ADL model:
  - same #lags for both $Y_{1,t}$ and $Y_{2,t}$ in both equations.
  - current value of additional explanatory variable ("X_t") is ruled out.

The Bivariate VAR Model

Recall the model:

$$\begin{aligned} Y_{1,t} &= \mu_1 + \sum_{k=1}^{p} \phi_{1,k} Y_{1,t-k} + \sum_{k=1}^{p} \gamma_{1,k} Y_{2,t-k} + \varepsilon_{1,t}, \\ Y_{2,t} &= \mu_2 + \sum_{k=1}^{p} \phi_{2,k} Y_{2,t-k} + \sum_{k=1}^{p} \gamma_{2,k} Y_{1,t-k} + \varepsilon_{2,t}. \end{aligned}$$

Collecting the coefficients in vectors and matrices,

$$Y_t = \begin{bmatrix} Y_{1,t} \\ Y_{2,t} \end{bmatrix}, \ \varepsilon_t = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}, \ \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \ \Phi_k = \begin{bmatrix} \phi_{1,k} & \gamma_{1,k} \\ \phi_{2,k} & \gamma_{2,k} \end{bmatrix}$$

for $k = 1, ..., p$, the above equations can be written as

$$Y_t = \mu + \Phi_1 Y_{t-1} + ... + \Phi_p Y_{t-p} + \varepsilon_t.$$

The VAR(p)-model is just a multivariate extension of the univariate AR(p) model.

Relationship with simultaneous equation structural models

- Consider the following model, where the errors are mutually uncorrelated:

$$\begin{aligned} Y_{1t} &= c_1 + d_1 Y_{1t-1} + \delta Y_{2t} + u_{1t} \\ Y_{2t} &= c_2 + \gamma Y_{1t} + d_2 Y_{2t-1} + u_{2t} \end{aligned}$$

- The model above is a simultaneous equation model (SEM). Models of this type are widely used in economics. Re-write as follows:

$$\begin{bmatrix} 1 & -\delta \\ -\gamma & 1 \end{bmatrix} \begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} \begin{bmatrix} Y_{1t-1} \\ Y_{2t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

$$A_0 Y_t = C + D Y_{t-1} + u_t$$

- No forecasts can be computed from this form, as we need the contemporaneous values of one variable to forecast the other.

Relationship with simultaneous equation structural models

- To forecast, we need the reduced form which provides us with the values of the variables as function of only shocks and past values.
- By premultiplying by A_0^{-1} we have:

$$Y_t = A_0^{-1} C + A_0^{-1} D Y_{t-1} + A_0^{-1} u_t = B_0 + B_1 Y_{t-1} + \varepsilon_t$$

where:

$$B_0 = \begin{bmatrix} B_0^{(1)} \\ B_0^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & -\delta \\ -\gamma & 1 \end{bmatrix}^{-1} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} B_1^{(11)} & B_1^{(12)} \\ B_1^{(21)} & B_1^{(22)} \end{bmatrix} = \begin{bmatrix} 1 & -\delta \\ -\gamma & 1 \end{bmatrix}^{-1} \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$$

$$\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} = \begin{bmatrix} 1 & -\delta \\ -\gamma & 1 \end{bmatrix}^{-1} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

- The reduced form is an unrestricted VAR

Structural VARs

- This means that, given a VAR:

$$Y_t = B_0 + B_1 Y_{t-1} + \varepsilon_t \tag{1}$$

- If we identify the shocks in some way, e.g. $\varepsilon_t = A_0^{-1} u_t$ where u_t has a diagonal variance matrix we can write:

$$A_0 Y_t = A_0 B_0 + A_0 B_1 Y_{t-1} + u_t \tag{2}$$

  which is a Structural VAR, i.e. a simultaneous equations model in which each structural shock u_t is uncorrelated with the others.

- The matrix A_0 describes the contemporaneous correlations across the variables.
  [Prior on B_0 or $A_0 B_0$?]
  [Uninformative sign restrictions on A_0^{-1} are uninformative for A_0?]
- The same reduced form (1) corresponds to several structural forms (2), so while it is easy to go from (2) to (1) the opposite is not obvious.

The General VAR Model

- The bivariate VAR model is easily extended to the general N-dimensional case.
- Let $Y_t = [Y_{1,t}, ..., Y_{k,t}]' \in \mathbb{R}^k$ be a N-dimensional vector of time series.
- The corresponding VAR(p) model is

$$Y_t = \mu + \Phi_1 Y_{t-1} + ... + \Phi_p Y_{t-p} + \varepsilon_t,$$

where $\mu \in \mathbb{R}^N$ is a vector, $\Phi_i \in \mathbb{R}^{N \times N}$ is a matrix, $i = 1, ..., p$, and $\varepsilon_t = [\varepsilon_{1,t}, ..., \varepsilon_{N,t}]' \in \mathbb{R}^N$ is a vector of errors.

- The errors are assumed to be i.i.d. with mean zero and covariance

$$\text{Var}(\varepsilon_t) = \Sigma$$

Stationarity and Characteristic polynomial

- **Characteristic polynomial:** Introducing the matrix polynomial corresponding to the VAR model,

$$\Phi(z) = I_N - \Phi_1 z - ... - \Phi_p z^p, \quad I_N \text{ is the identity matrix,}$$

we can write the VAR(p) model as

$$\Phi(L) Y_t = \mu + \varepsilon_t.$$

- **Stationarity:** Suppose $\{\varepsilon_t\}$ is i.id. $(0, \Sigma)$. Then the VAR(p) process is stationary and mixing if all the roots of $\Phi(z)$ lie outside the unit circle:

$$|\Phi(z)| = 0 \Rightarrow |z| > 1.$$

Moreover, $E\left[\|Y_t\|^p\right] < \infty$ if $E\left[\|\varepsilon_t\|^p\right] < \infty$.

ACF of a stationary VAR(1)

$$Y_t = \mu + \Phi Y_{t-1} + \varepsilon_t.$$

First and second moments:

$$\begin{aligned} E[Y_t] &= \mu + \Phi E[Y_{t-1}] \rightarrow \mu_Y = (I - \Phi)^{-1} \mu \\ Var[Y_t] &= \Phi Var[Y_{t-1}]\Phi' + \Sigma \rightarrow V_Y = \Phi V_Y \Phi' + \Sigma \\ vec(Var[Y_t]) &= (I - \Phi \otimes \Phi')^{-1} vec(\Sigma) \end{aligned}$$

Autocovariance function:

$$\begin{aligned} Cov[Y_t, Y_{t-1}] &= Cov[\Phi Y_{t-1} + \varepsilon_t, Y_{t-1}] = \Phi V_Y \\ Cov[Y_t, Y_{t-2}] &= Cov[\Phi Y_{t-1} + \varepsilon_t, Y_{t-2}] = \Phi Cov[Y_t, Y_{t-1}] \\ &= \Phi^2 V_Y \ ... \\ Cov[Y_t, Y_{t-k}] &= Cov[\Phi Y_{t-1} + \varepsilon_t, Y_{t-k}] = \Phi^k V_Y \\ Corr\,(Y_t, Y_{t-k}) &= \Phi^k \end{aligned}$$

Forecasting with a VAR

- Consider for example a VAR(1) $Y_t = \mu + \Phi Y_{t-1} + \varepsilon_t$. Its MSFE-optimal forecast is:

$$\hat{Y}_{t+1} = E(Y_{t+1}|I_t) = \mu + \Phi Y_t$$

- What happens if we go further ahead in the forecasting horizon?

$$\hat{Y}_{t+2} = \mu + \Phi \hat{Y}_{t+1} = \mu + \Phi(\mu + \Phi Y_t) = \mu + \Phi\mu + \Phi^2 Y_t$$

- IMPORTANT: the notation Φ^2 means $\Phi = \Phi\Phi$, which is different from squaring the elements of Φ!
- The h-step ahead forecast is:

$$\hat{Y}_{t+h} = \mu + \Phi\mu + \Phi^2\mu + ... + \Phi^{h-1}\mu + \Phi^h Y_t$$

- If the horizon is infinite, we have:

$$\hat{Y}_{t+\infty} = (\Sigma_{i=1}^{\infty}\Phi)\mu = (I - \Phi)^{-1}\mu = E[Y_t]$$

Forecast errors

- The 1-step ahead forecast error is:

$$\begin{aligned}
\mathbf{v}_{t+1} &= Y_{t+1} - E(Y_{t+1}|I_t) \\
&= Y_{t+1} - (\mu + \Phi Y_t) \\
&= (\mu + \Phi Y_t + \varepsilon_{t+1}) - (\mu + \Phi Y_t) = \varepsilon_{t+1}
\end{aligned}$$

- The 2-step ahead forecast error is:

$$\begin{aligned}
\mathbf{v}_{t+2} &= Y_{t+2} - E(Y_{t+2}|I_t) \\
&= (\mu + \Phi Y_{t+1} + \varepsilon_{t+2}) - (\mu + \Phi\mu + \Phi Y_t) \\
&= \Phi Y_{t+1} + \varepsilon_{t+2} - \Phi\mu - \Phi Y_t \\
&= \Phi(Y_{t+1} - (\mu + \Phi Y_t)) + \varepsilon_{t+2} \\
&= \Phi\varepsilon_{t+1} + \varepsilon_{t+2}
\end{aligned}$$

Forecast errors

- The h-step ahead forecast error is:

$$\begin{aligned} \mathbf{v}_{t+h} &= Y_{t+h} - E(Y_{t+h}|I_t) \\ &= \Phi^{h-1}\varepsilon_{t+1} + ... + \Phi\varepsilon_{t+h-1} + \varepsilon_{t+h} = \sum_{j=1}^{h} \Phi^{h-j}\varepsilon_{t+j} \end{aligned}$$

- The variance of the h-step ahead forecast error is:

$$Var(\mathbf{v}_{t+h}) = \Phi^{h-1}\Sigma\Phi^{h-1\prime} + ... + \Phi\Sigma\Phi' + \Sigma$$

where Σ is the variance (matrix) of the error term ε_t.

- As in the univariate case, the forecast errors are correlated with correlation:

$$\begin{aligned} Cov(\mathbf{v}_{t+2}, \mathbf{v}_{t+1}) &= Cov(\Phi\varepsilon_{t+1} + \varepsilon_{t+2}, \varepsilon_{t+1}) = \Phi\Sigma \\ Cov(\mathbf{v}_{t+3}, \mathbf{v}_{t+1}) &= Cov(\Phi^2\varepsilon_{t+1} + \Phi\varepsilon_{t+2} + \varepsilon_{t+3}, \varepsilon_{t+1}) = \Phi^2\Sigma \\ Cov(\mathbf{v}_{t+h}, \mathbf{v}_{t+1}) &= Cov(\Phi^{h-j}\varepsilon_{t+j} + ..., \varepsilon_{t+1,}) = \Phi^{h-j}\Sigma \end{aligned}$$

Variance decomposition

- The expression

$$Var(\mathbf{v}_{t+h}) = \Phi^{h-1}\Sigma\Phi^{h-1\prime} + ... + \Phi\Sigma\Phi' + \Sigma$$

  is particularly useful because it can be use to decompose the total variance of the innovations $\mathbf{v}_{t+h}$ in the contributions given by each of the variables in the VAR.
- However, in order to do so in an economically meaningful way **one needs to "rotate" the errors ε_t so that they are orthogonal**
- The simplest (but by no means the only) way to do so is by defining:

$$\varepsilon_t = A_0^{-1} u_t, \text{ with } Var(u_t) = \Lambda,$$

  with A_0^{-1} lower triangular with ones on the main diagonal, and Λ diagonal. By construction we have:

$$\Sigma = A_0^{-1}\Lambda A_0^{-1\prime}$$

MA Representation of Stationary VARs

- Example with VAR(1):

$$
\begin{aligned}
Y_t &= \mu + \Phi_1 Y_{t-1} + \varepsilon_t \\
(I - \Phi_1 L) Y_t &= \mu + \varepsilon_t \text{ (AR representation, } \Phi(L) = (I - \Phi_1 L)) \\
Y_t &= \mu + \Phi_1 Y_{t-1} + \varepsilon_t \\
&= \mu + \Phi_1 (\mu + \Phi_1 Y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\
&= \mu + \Phi_1 (\mu + \Phi_1 (\mu + \Phi_1 Y_{t-3} + \varepsilon_{t-2}) + \varepsilon_{t-1}) + \varepsilon_t \\
&.... \\
&= \mu (I + \Phi_1 + \Phi_1^2 + ... + \Phi_1^\infty) \\
&\quad + (I + \Phi_1 L + \Phi_1^2 L^2 + ... + \Phi_1^\infty L^\infty) \varepsilon_t \\
&= \mu \Pi(1) + \Pi(L) \varepsilon_t \text{ (MA, } \Pi(L) = (I - \Phi_1 L)^{-1}) \\
\Pi(1) &\rightarrow (I - \Phi_1)^{-1} \text{ and } \mu (I - \Phi_1)^{-1} = \mu_Y \text{, then} \\
Y_t &= \mu (I - \Phi_1)^{-1} + \Pi(L) \varepsilon_t = \mu_Y + \Pi(L) \varepsilon_t
\end{aligned}
$$

MA Representation and impulse responses.

- The MA representation can be used to compute impulse responses:

$$\Phi(L) Y_t = \mu + \varepsilon_t \Leftrightarrow Y_t = \mu_Y + \Pi(L)\varepsilon_t$$

- Define the notation $\Pi(L) = I + \Pi_0 + \Pi_1 L + ... + \Pi_h L^h + ...$, where Π_l has generic element $[\Pi_l]_{i,j}$. We have:

$$\frac{\partial Y_{i,t+h}}{\partial \varepsilon_{j,t}} = [\Pi_l]_{i,j}$$

- As with the variance decomposition, there is not too much economic meaning unless we rotate the disturbances. Defining again

$$\varepsilon_t = A_0^{-1} u_t, \text{ with } Var(u_t) = \Lambda,$$

allows to compute instead:

$$\frac{\partial Y_{i,t+h}}{\partial u_{j,t}} = [\Pi_l^*]_{i,j}$$

where

$$Y_t = \mu_Y + \Pi(L)A_0^{-1} u_t = \mu_Y + \Pi^*(L) u_t$$

VARs - Multivariate regression representation

- Consider the following VAR:

$$y_t = B_0 + B_1 y_{t-1} + B_2 y_{t-2} + ... + B_p y_{t-p} + u_t$$

- By collecting $B = (B_0,\ B_1,\ ..., B_p)'$ of dimension $N \times k$ (where $k = 1 + Np$) and $x_t = (1, y'_{t-1}, y'_{t-2}, ..., y'_{t-p})'$ of dimension $k \times 1$ we have:

$$y_t = B' x_t + v_t.$$

- Now consider the equations for all observations $t = 1, ..., T$. By stacking them by columns and then transposing the system we get the multivariate regression:

$$\underset{T\times N}{Y} = \underset{T\times k}{X}\ \underset{k\times N}{B} + \underset{T\times N}{V} \qquad V \sim N(0, \underset{N\times N}{\Sigma})$$

  where Y is a data-matrix with rows y'_t, X is a data-matrix with rows $x'_t = (1, y'_{t-1}, y'_{t-2}, ..., y'_{t-p})$ and V is a data-matrix with rows v'_t.
- The FIML estimator coincides with OLS and is $\hat{B} = (X'X)^{-1}X'Y$

VARs - Vectorized representation

- Vectorizing:

$$\underset{T\times N}{vec(Y)} = \underset{T\times k}{vec(X}\ \underset{k\times N}{B)} + \underset{T\times N}{vec(V)} \qquad V \sim N(0, \underset{N\times N}{\Sigma})$$

gives:

$$y = (I_N \otimes X)\beta + v \qquad v \sim N(0, \Sigma \otimes I_T)$$

- The OLS estimator is:

$$\hat{\beta} = (I_N \otimes (X'X)^{-1}X')y = vec((X'X)^{-1}X'Y)$$

that is, equation by equation is equivalent to FIML system estimation.

- This happens because of the peculiar structure of $\Sigma \otimes I_T$
- The VAR is a special case of a SUR model in which all the regressors are the same. If we impose some restrictions on the coefficients of the VAR then it becomes a SUR model in which the regressors in each equation are different, hence FIML or 3SLS would be required.

Inference based on OLS estimator

- If the VAR model is stable, OLS estimator is normally distributed *in large samples*.
- It is consistent (but only if the VAR has sufficient lags to ensure the error is a MDS)
- Testing simple hypotheses: t-stat's are standard normally dist'd.
- Testing joint hypotheses: F-stat's follow F-dist's.
- **Multiple equations and F-test:** You may want to test hypotheses involving several equations. For example, $H_0 : B_p = 0$ in $Y_t = \sum_{i=1}^{p} B_i Y_{t-i} + \varepsilon_t$. You can use F- or LR-statistics for this since the OLS estimators across equations are jointly normally distributed.

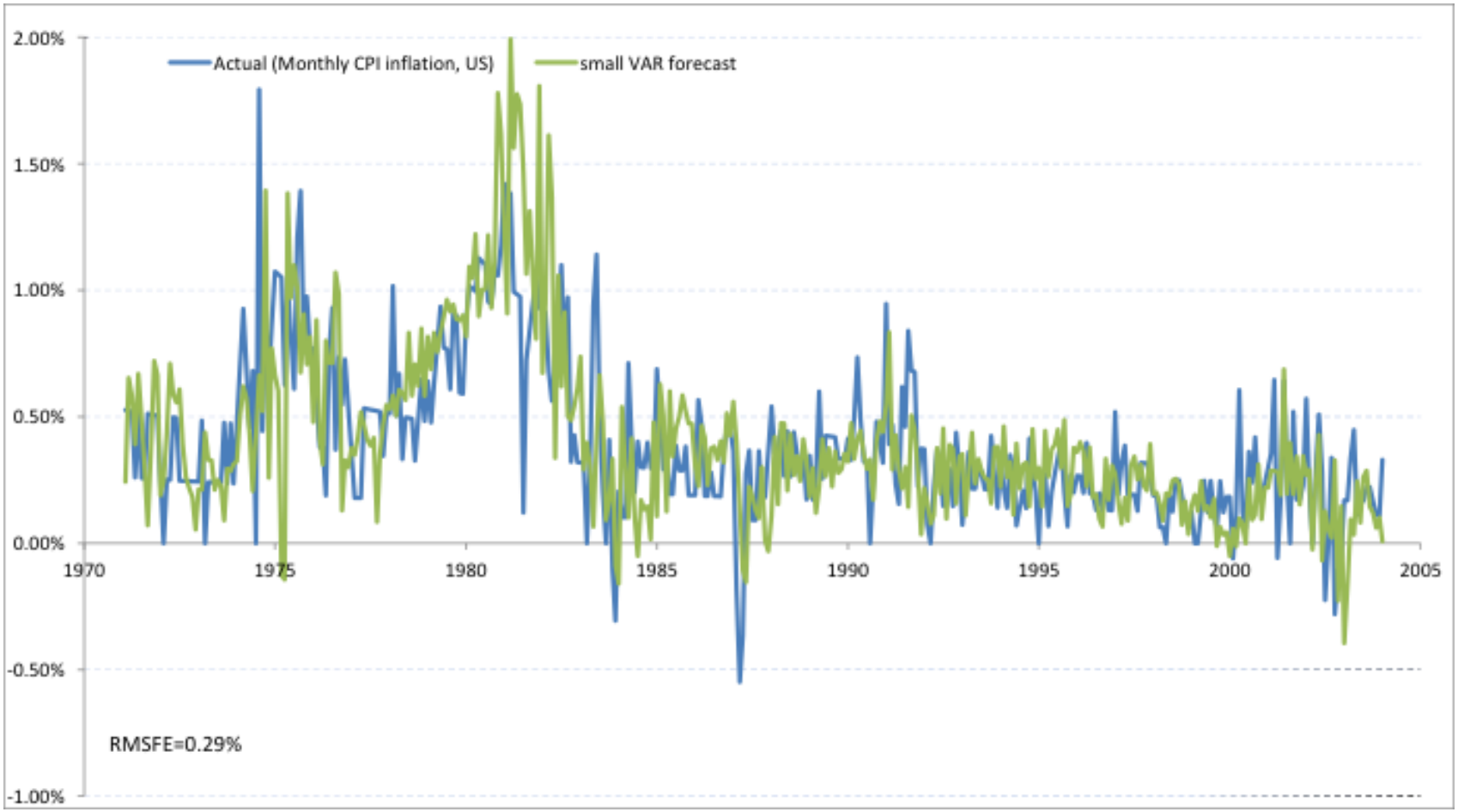
Inference based on OLS estimator

- Some things that are much harder to do on the basis of OLS estimates are:
  - Impulse responses
  - Variance decompositions
  - Forecasts (for h$>$1, both point and density)
- This is because all the functions above are nonlinear functions of the VAR coefficients!
- For proper inference, we need to use the delta method, or bootstrap

Bayesian VARs

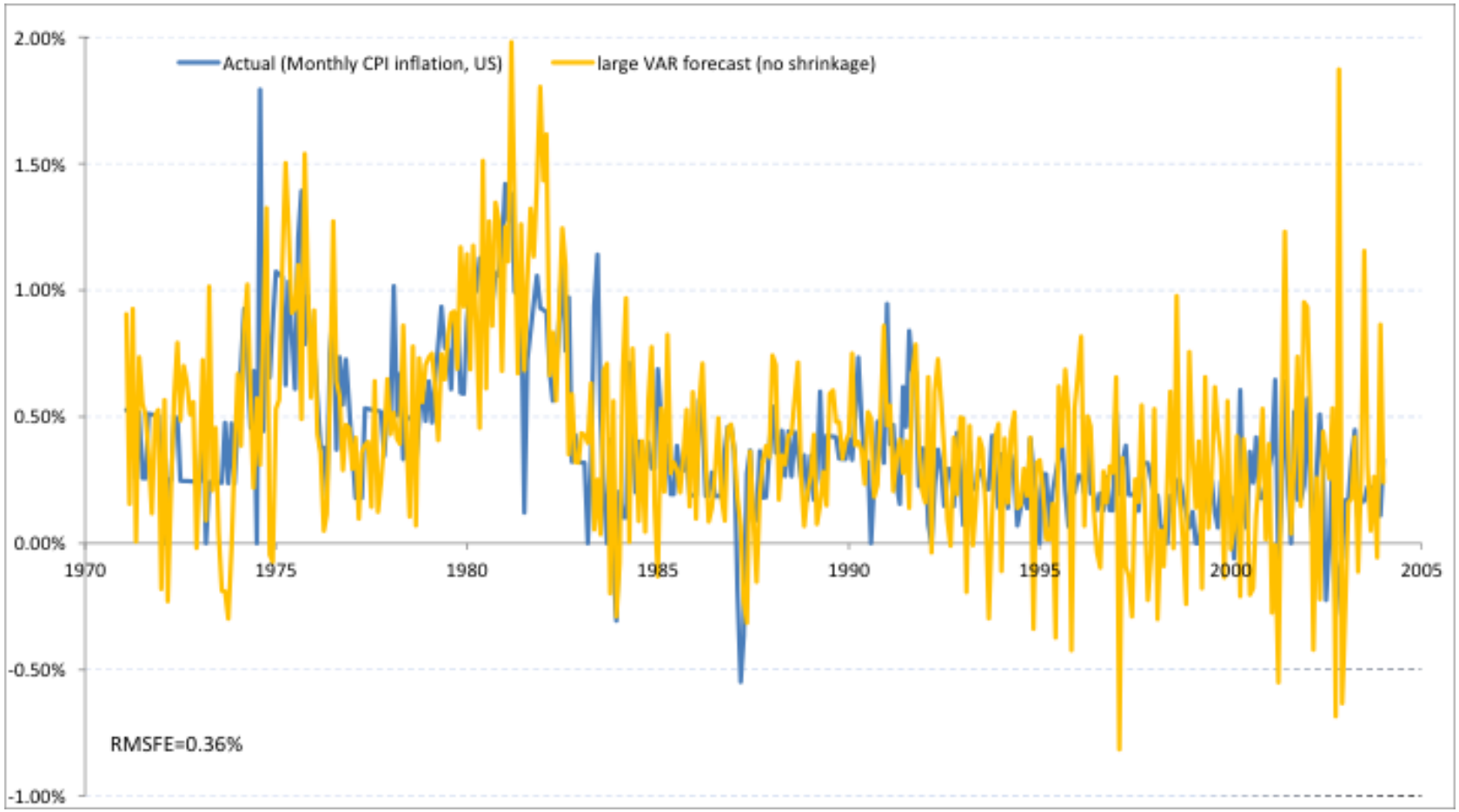
- The outstanding ability of Bayesian methods in forecasting has been known since the works of Litterman (1979) and Doan, Litterman, and Sims (1984)
- Bayesian VARs offer three main advantages.
  1. They are particularly well suited in handling very large cross-sections of data, even when the time series available are short.
  2. They offer a theoretically grounded way to impose judgmental information and a-priori beliefs in the model.
  3. They provide a natural environment to produce forecasts of the whole distribution of a time series, i.e. fan charts.

Bayesian VARs



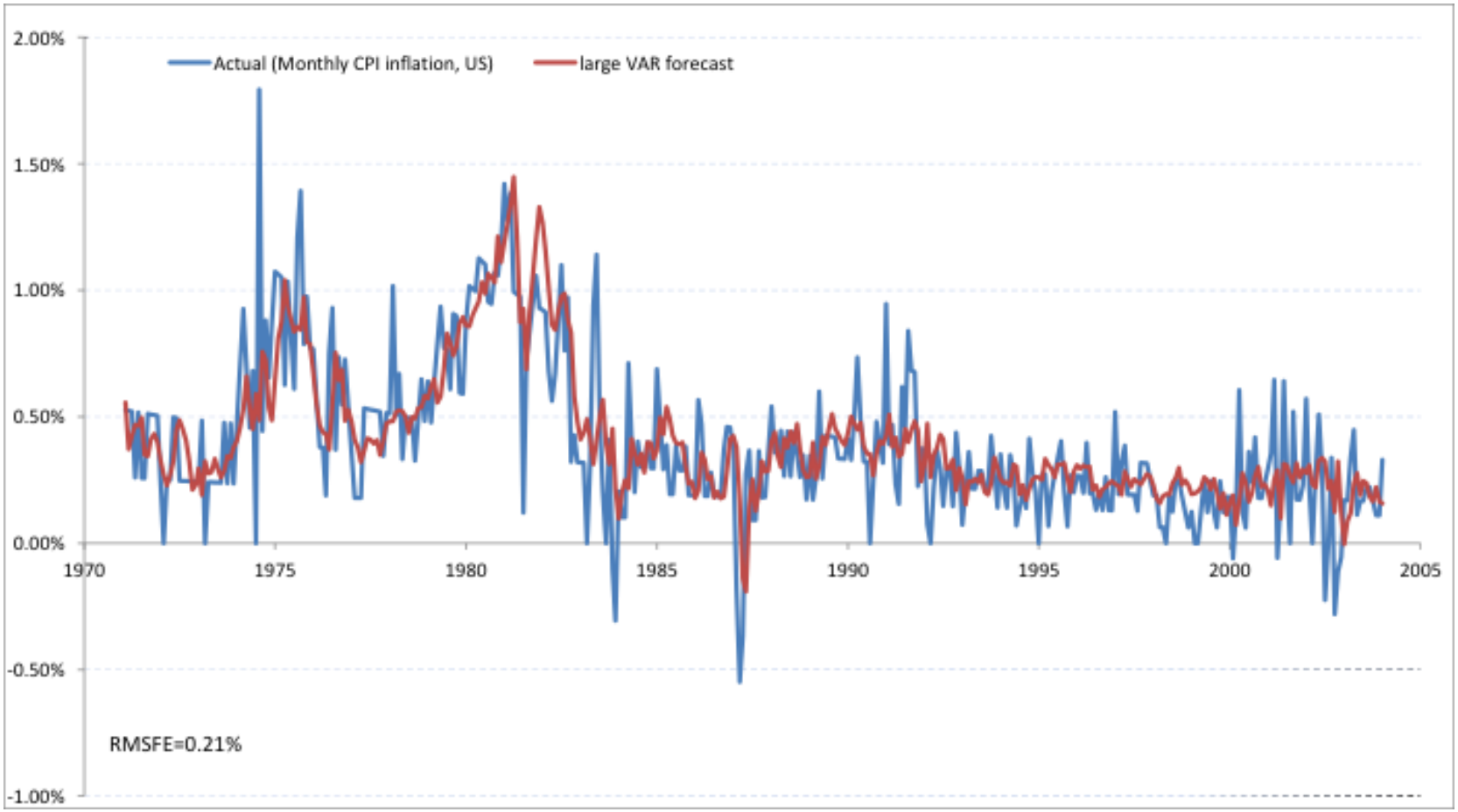
Small VAR

Bayesian VARs



Large VAR

Bayesian VARs



Large VAR with shrinkage

Bayesian VARs

- In a Bayesian VAR, the matrix of coefficients B is random.
- One can specify his/her beliefs on the values of B as follows:

$$\beta \sim N(\beta_0, \Omega_0);$$

- The vector β_0 is the prior mean. One can set it to the values he/she believes in
- The matrix Ω_0 is the variance around β_0. It measures how uncertain we are about our prior beliefs.

Bayesian VARs

We will now proceed to study the same 5 cases we considered for the univariate model.

- Fix the error variance with an estimate (Theil's mixed estimator)
- The Natural Conjugate N-IG prior
- The Independent N-IG prior
- The diffuse (Jeffrey's) prior
- The Normal-diffuse (Zellner's) prior

Theil mixed estimator

The simplest case is the one in which the error variance is estimated in a preliminary step and treated as known. Combining the prior $\beta \sim N(\beta_0, \Omega_0)$ with the likelihood:

$$p(y|\beta, \hat{\Sigma}) = (2\pi)^{-\frac{TNk}{2}} |\hat{\Sigma} \otimes I_T|^{-\frac{1}{2}} \times \exp[-(y - (I_N \otimes X)\beta)' (\hat{\Sigma} \otimes I_T)^{-1} (y - (I_N \otimes X)\beta)/2]$$

The conditional posterior kernel is:

$$\begin{aligned} p(\beta|y, \hat{\Sigma}) &\propto p(y|\beta) \times p(\beta) \propto \exp[-(\beta - \beta_0)' \Omega_0^{-1} (\beta - \beta_0)/2] \\ &\quad \times \exp[-(y - (I_N \otimes X)\beta)' (\hat{\Sigma} \otimes I_T)^{-1} (y - (I_N \otimes X)\beta)/2] \\ &\propto \exp[-(\beta - \beta_1)' \Omega_1^{-1} (\beta - \beta_1)/2] \end{aligned}$$

with

$$\Omega_1^{-1} = \Omega_0^{-1} + (I_N \otimes X)'(\hat{\Sigma} \otimes I_T)^{-1}(I_N \otimes X),$$

$$\beta_1 = \Omega_1 \left(\Omega_0^{-1} \beta_0 + (I_N \otimes X)'(\hat{\Sigma} \otimes I_T)^{-1} y \right).$$

Theil mixed estimator

- This yields:

$$\beta|y,\Sigma \sim N(\beta_1,\Omega_1)$$

with:

$$\Omega_1^{-1} = \Omega_0^{-1} + (\hat{\Sigma}^{-1} \otimes X'X) = \Omega_0^{-1} + \left(\sum_{t=1}^{T} X_t'\hat{\Sigma}^{-1}X_t\right)^{-1}$$

$$\beta_1 = \Omega_1\left(\Omega_0^{-1}\beta_0 + (\hat{\Sigma}^{-1} \otimes X')y\right) = \Omega_1\left(\Omega_0^{-1}\beta_0 + \left(\sum_{t=1}^{T} X_t'\hat{\Sigma}^{-1}y_t\right)^{-1}\right)$$

this is the conditional distribution of β under the independent N-IW prior.

- As $\Omega_0^{-1} \longrightarrow 0$ the prior information becomes irrelevant:

$$\beta_1 \longrightarrow (\hat{\Sigma}^{-1} \otimes X'X)^{-1}(\hat{\Sigma}^{-1} \otimes X')y = (I_N \otimes (X'X)^{-1}X')y = \hat{\beta}$$

The Minnesota (Litterman's) prior

- The Minnesota prior in its original implementation is of the Theil mixed estimation form
- Then, the Minnesota prior moments can also be used to specify the moments of other form of priors (N-IW conjugate and not, Normal-diffuse).
- A-priori each variable in the VAR follows a RW $\mathbf{y}_t = 0 + I\mathbf{y}_{t-1} + 0\mathbf{y}_{t-2} + ... + 0\mathbf{y}_{t-p} + \varepsilon_t$. That is, for $k = 0, ..., p$,:

$$E[B_k^{(ij)}] = 1 \text{ if } i = j,\ k = 1;\ E[B_k^{(ij)}] = 0 \text{ otherwise.}$$

- The uncertainty around such prior mean is given by:

$$Var[B_k^{(ij)}] = \lambda_1 \times \lambda_2(1_{i \neq j}) \times \frac{1}{k^{\lambda_3}} \times \sigma_i^2/\sigma_j^2,\ k = 1, ..., p;$$

- λ_1 measures the tightness of the prior: when $\lambda_1 \to 0$ the prior is imposed exactly, while as $\lambda_1 \to \infty$ estimates will approach the *OLS* estimates.
- λ_2 controls the standard deviation of the prior on lags of variables other than the dependent variable. With $\lambda_2 = 1$ there is no distinction between lags of the dependent variable and other variables.
- λ_3 controls the decay over lags

The Minnesota prior

- $\frac{\sigma_i}{\sigma_j}$ are scaling parameters. These are estimated from univariate AR regressions, $\hat{\sigma}_i^2 / \hat{\sigma}_j^2$
- The error variance is estimated in a preliminary step using $\hat{\sigma}_i^2$ and assumed diagonal.
- Usually the suggested values for the hyperparameters are:

$$\lambda_1 = 0.2;\ \lambda_2 = 1;\ \lambda_3 = 1 \textit{ or } 2;$$

  but there are ways to choose these optimally (See e.g. Carriero, Clark, Marcellino 2012 and Giannone, Lenza, Primiceri 2012)
- The forecast will be then a weighted average of an OLS and a RW.
- It works remarkably well in macroeconomic applications
- Marginal likelihood available in closed form
- One can also impose cointegration and unit roots (sharply or as a prior)

The Natural conjugate case

- An alternative prior is the natural-conjugate:

$$\beta|y,\Sigma \sim N(\beta_0,\Omega_0);\ \Sigma \sim IW(S_0,\nu_0)$$

  where $vec(B_0) = \beta_0$ and where:

$$\Omega_0 = \Sigma \otimes \Psi_0$$

- Note that the prior for β is specified conditionally on the knowledge of Σ.

Wishart and Inverse Wishart

- We have that
$$\Sigma^{-1} \sim W(S_0^{-1}, \nu_0) \Leftrightarrow \Sigma \sim IW(S_0, \nu_0)$$
- The Wishart pdf is:
$$p(\Sigma^{-1}) \propto |S_0^{-1}|^{-\nu_0/2} |\Sigma^{-1}|^{-(\nu_0 - N - 1)/2} \exp\{-0.5 tr(\Sigma S_0^{-1})\}$$
with mean $E[\Sigma^{-1}] = \nu_0 S_0^{-1}$
- The Inverse Wishart pdf is:
$$p(\Sigma) \propto |S_0|^{\nu_0/2} |\Sigma|^{-(\nu_0 - N - 1)/2} \exp\{-0.5 tr(\Sigma^{-1} S_0)\}$$
The mean $E[\Sigma] = \frac{1}{\nu_0 - N - 1} S_0$
- This is simply the multivariate version of an inverse gamma

Drawing from an Inverse Wishart

- The "notional data" interpretation of this prior distribution is the information about precision from ν_0 i.i.d. N-variate normal observations with sum of squares S_0.
- To draw Σ we can:
  1. Draw a **matrix** $A = \underset{N\times N}{S_0^{-1/2}}\ \underset{N\times\nu_0}{v_{1:\nu_0}}$ of ν_0 random vectors from $A \sim N(0, I_N)$;
  2. The quantity $(AA') = v' S_0^{-1} v$ is a random draw from $W(v_0, S_0)$
  3. $(AA')^{-1} = (v' S_0^{-1} v)^{-1}$ is a draw from the corresponding Inverse Wishart $IW(v_0, S_0)$.

Matricvariate Normal

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

$$Z \sim MN(M, Q, P)$$

where M is $p \times q$ and P and Q are positive definite symmetric matrices of dimensions $p \times p$ and $q \times q$ if $x = vec(X)$ is multivariate normal:

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

- The density is:

$$\begin{aligned} p(X) &= 2\pi^{-pq/2} |Q \otimes P|^{-1/2} \\ &\times \exp\left[-\frac{1}{2}(z - vec(M))'(Q \otimes P)^{-1}(z - vec(M))\right] \\ = \quad & 2\pi^{-pq/2} |Q|^{-p/2} |P|^{-q/2} \exp\left[-\frac{1}{2} tr\left\{Q^{-1}(Z - M)'P^{-1}(Z - M)\right\}\right] \end{aligned}$$

where we used $tr(ABCD) = vec(A')'(D' \otimes B)vec(C)$.

Multivariate regression likelihood

In the case of the multivariate regression we have that:

$$\underset{T\times N}{Y} = \underset{T\times k}{X}\ \underset{k\times N}{B} + \underset{T\times N}{V} \qquad V \sim MN(0, \underset{N\times N}{\Sigma}, I_T)$$

with likelihood:

$$\begin{aligned} p(B,\Sigma|Y) &= 2\pi^{-pq/2}\,|\Sigma|^{-T/2}\,|I_k|^{-k/2} \\ &\quad \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-XB)'(Y-XB)\right\}\right] \\ &\propto |\Sigma|^{-(T-k)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \\ &\quad |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-\hat{B})'(B-\hat{B})\right\}\right] \end{aligned}$$

which is in the form of an $\Sigma \sim IW(\hat{S}, T-k)$ times a matricvariate normal for B conditional on Σ

The Natural conjugate case

This prior yields the following posterior kernel:

$$
\begin{aligned}
p(\beta,\sigma^2|Y) \quad \propto \quad & |\Sigma|^{-(T-k)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\hat{S}\right\}\right] \\
& |\Sigma|^{-(v_0+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}S_0\right\}\right] \\
& |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-\hat{B})'(B-\hat{B})\right\}\right] \\
& |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-B_0)'\Psi_0^{-1}(B-B_0)\right\}\right] \\
= \quad & |\Sigma|^{-(v_1+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}S_1\right\}\right] \\
& |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-B_1)'\Psi_1^{-1}(B-B_1)\right\}\right]
\end{aligned}
$$

The Natural conjugate case

- That is

$$B|Y,\Sigma \sim MN(B_1,\Sigma,\Psi_1);\ \Sigma \sim IW(S_1,\nu_1)$$

with:

$$\Omega_1 = \Sigma \otimes \Psi_1$$

$$\begin{aligned}
\nu_1 &= \nu_0 + T, \\
S_1 &= S_0 + Y'Y + B_0'\Psi_0^{-1}B_0 - B_1'\Psi_1^{-1}B_1 \\
\Psi_1 &= (\Psi_0^{-1} + X'X)^{-1}, \\
B_1 &= \Psi_1(\Psi_0^{-1}B_0 + X'Y)
\end{aligned}$$

- Equivalently for $vec(B_1) = \beta_1$:

$$\beta|y,\Sigma \sim N(\beta_1,\Sigma \otimes \Psi_1);\ \Sigma \sim IW(S_1,\nu_1)$$

The Natural conjugate case : symmetry

This posterior kernel can also be written in the multivariate normal form:

$$
\begin{aligned}
& p(\beta, \sigma^2 | Y) \\
\propto \quad & |\Sigma|^{-\frac{T+\nu_0-1-N}{2}} \exp\left[-\frac{1}{2} tr(S_0 \Sigma^{-1})\right] \\
& |\Sigma|^{-\frac{1}{2}} \exp\left[\begin{array}{c} -\frac{1}{2}(y-(I_N \otimes X)\beta)'(\Sigma \otimes I_T)^{-1}(y-(I_N \otimes X)\beta) \\ -\frac{1}{2}(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0) \end{array}\right]
\end{aligned}
$$

which has a matricvariate representation because $\Omega_0^{-1} = \Sigma^{-1} \otimes \Psi_0^{-1}$ and $var(v) = \Sigma \otimes I_T$.

The Natural conjugate prior : symmetry

Note that the expression for the posterior variance can be obtained using the formulas we used before, and exploiting the Kronecker structure:

$$\begin{aligned}
\Omega_1^{-1} &= \Omega_0^{-1} + (I_N \otimes X)'(\Sigma \otimes I_T)^{-1}(I_N \otimes X) \\
&= \underbrace{(\Sigma^{-1} \otimes \Psi_0^{-1})}_{Symmetric\ Prior} + \underbrace{(\Sigma^{-1} \otimes X'X)}_{Symmetric\ likelihood} \\
&= \Sigma^{-1} \otimes (\Psi_0^{-1} + X'X) = \Sigma^{-1} \otimes \Psi_1^{-1}
\end{aligned}$$

However, to do this we need to assume Ω_0^{-1} has a Kronecker structure and is specified under knowledge of Σ!

Advantages of the natural conjugate prior

- Advantages:
  1. Simple MC sampling: $p(\beta|\Sigma, y)p(\Sigma|y) = p(\beta, \Sigma|y)$
  2. Computational complexity of order N^3 rather than N^6 $\rightarrow$use $chol(\Psi_1) * \text{rand}(k, N) * chol(\Sigma)'$
  3. Marginal likelihood exists in closed form: $p(y) = \frac{[p(\beta|\Sigma)p(\Sigma)] \times p(y|\beta,\Sigma)}{p(\beta|\Sigma,y)p(\Sigma|y)}$
  4. Implementable with dummy variables
- Shortcoming: no asymmetric priors are allowed (also, no asymmetric likelihoods, e.g. no stochastic volatility)

Monte Carlo sampling

- Simple MC sampling from the joint posterior $p(\beta, \Sigma|y)$
  - Draw $p(\Sigma|y)$
  - Draw $p(\beta|\Sigma, y)$
- To draw Σ we can:
  1. Draw a **matrix** $A = \underset{N\times N}{S_0^{-1/2}} \underset{N\times\nu_0}{v_{1:\nu_0}}$ of ν_0 random vectors from $A \sim N(0, I_N)$;
  2. The quantity $(AA') = v'S_0^{-1}v$ is a random draw from $W(v_0, S_0)$
  3. $(AA')^{-1} = (v'S_0^{-1}v)^{-1}$ is a draw from the corresponding Inverse Wishart $IW(v_0, S_0)$.

Computational gains

- Drawing a sequence of β can be in general rather demanding, but in this case the matricvariate structure considerably speeds up the computations.
- An intuitive way to draw β, conditionally on a draw of the error variance Σ, is:

$$vec(B) = vec(B_1) + chol(\Sigma \otimes \Psi_1) \times v \tag{3}$$

  where v is a $kN \times 1$ standard Gaussian vector process.
- The Choleski decomposition above requires $(kN)^3$ elementary operations.
- However by organizing the elements of v in a $k \times N$ matrix V such that $v = vec(V)$, one could draw the matrix Φ as follows:

$$\Phi = B_1 + chol(\Psi_1) \times V \times chol(\Sigma)'. \tag{4}$$

- This speeds up the computations by a factor of N^3, because the two Choleski decompositions $chol(\bar{\Omega})$ and $chol(\Sigma)$ require only $k^3 + N^3$ operations

Marginal Likelihood

The marginal likelihood is matricvariate t:

$$y = (I_N \otimes X)\beta + \varepsilon$$

with $\varepsilon \sim N(0, \Sigma \otimes I)$. Since $\beta|\Sigma \sim N(\beta_1, \Sigma \otimes \Psi_0)$ then

$$(I_N \otimes X)\beta|\Sigma \sim N((I_N \otimes X)\beta_0, (I_N \otimes X)(\Sigma \otimes \Psi_1)(I_N \otimes X')).$$

It follows that

$$\begin{aligned} y|\Sigma &\sim N((I_N \otimes X)\beta_0, (I_N \otimes X)(\Sigma \otimes \Psi_0)(I_N \otimes X') + (\Sigma \otimes I)) \\ &= N((I_N \otimes X)\beta_0, \Sigma \otimes (X\Psi_0 X' + \otimes I)) \end{aligned}$$

because ε and β are independent when conditioning on σ^2.

Marginal Likelihood

This is a normal, and Σ an inverse Wishart, so integrating this out gives a t:

$$y \sim t((I_N \otimes X)\beta_0, (X\Psi_0 X' + I), S_0, v_0)$$

which has pdf:

$$p(Y) = \pi^{\frac{-TN}{2}} \times |(I + X\Omega_0 X')^{-1}|^{\frac{N}{2}} \times |S_0|^{\frac{v_0}{2}} \times \frac{\Gamma_N(\frac{v_0+T}{2})}{\Gamma_N(\frac{v_0}{2})}$$
$$\times |S_0 + (Y - XB_0)'(I + X\Omega_0 X')^{-1}(Y - XB_0)|^{-\frac{v_0+T}{2}}, \quad (5)$$

derivation based on theorem A.19 in Bauwens, Lubrano and Richard (1999)

Marginal Likelihood

- Since

$$S_0 + (Y - XB_0)'(I + X\Omega_0 X')^{-1}(Y - XB_0) = S_1,$$

and

$$\left|X\Omega_0 X' + I\right| = |\Omega_0|\,|\Omega_1|^{-1}$$

this gives:

$$p(Y) = \pi^{\frac{-TN}{2}} \times \frac{\Gamma_N\left(\frac{v_0+T}{2}\right)}{\Gamma_N\left(\frac{v_0}{2}\right)} \times \frac{|\Omega_1|^{-N/2}}{|\Omega_0|^{-N/2}} \frac{|S_1|^{-\frac{v_0+T}{2}}}{|S_0|^{-v_0/2}} \tag{6}$$

- A similar expression can be obtained for the Litterman prior (fixed variance matrix).

Hyperparameters (tightness)

- The value of the marginal likelihood in (6) is provided by default in some computer packages such as Eviews.
- Del Negro and Schorfheide (2004), Carriero, Kapetanios, and Marcellino (2012): choose prior tightness by maximizing the marginal data density of the model
- Giannone, Lenza, Primiceri (2016): treat tightness as a coefficient and estimate it.

How to elicit the prior moments

- How to specify the prior moments?
- Use very well known stylized fact on macroeconomic time series. Litterman (1979) and Doan, Litterman, and Sims (1984)
- Use beliefs about the long run values of the variables (Villani 2011)
- Use economic or finance theory (Ingrahm and Whiteman 1989, Del Negro and Schorfheide 2004, Carriero 2015)

Dummy variable implementation of priors

- **Method 1 (fixed variance)**.
  Model $y = (I_N \otimes X)\beta + v \qquad v \sim N(0, \hat{\Sigma} \otimes I_T)$. We believe that $\beta \sim N(\beta_0, \Omega_0)$, which can be written as:

$$-u = (\beta - \beta_0) \sim N(0, \Omega_0) \rightarrow \beta_0 = \beta + u$$

  and appended to the system:

$$\begin{bmatrix} \overset{y^*}{\beta_0} \\ y \end{bmatrix} = \begin{bmatrix} \overset{Z^*}{I} \\ (I_N \otimes X) \end{bmatrix} \beta + \begin{bmatrix} u \\ v \end{bmatrix}; \quad Var\left(\begin{bmatrix} u \\ v \end{bmatrix}\right) = \begin{bmatrix} \overset{\Omega^*}{\Omega_0} & 0 \\ 0 & \hat{\Sigma} \otimes I_T \end{bmatrix}$$

- This system can be estimated with GLS (Theil 1971). The GLS estimator $\bar{b}$ is:

$$\bar{b} = \left(Z'^* \Omega^{*-1} Z^*\right)^{-1} \left(Z'^* \Omega^{*-1} y^*\right)$$

- Using $\Omega = \Sigma \otimes I_T$ we can get back to the formula given before:

$$\begin{aligned} \bar{b} &= [\Omega_0^{-1} + (I_N \otimes X)'(\hat{\Sigma} \otimes I_T)^{-1}(I_N \otimes X)]^{-1} \\ & (\Omega_0^{-1}\beta_0 + (I_N \otimes X)'(\hat{\Sigma} \otimes I_T)^{-1} y) = \beta_1 \end{aligned}$$

Dummy variable implementation of priors

- **Method 2 (random variance)** (e.g. Sims and Zha)
- Consider the conjugate N-IW prior, and assume we can write it as follows:

$$\begin{aligned} \Omega_0 &= \left(X_D' X_D\right)^{-1}; B_0 = \left(X_D' X_D\right)^{-1} \left(X_D' Y_D\right) \\ S_0 &= \left(Y_D - X_D B_0\right)' \left(Y_D - X_D B_0\right); \; v_0 = T_D \end{aligned}$$

  where X_D is a $T_D \times k$ matrix and Y_D is a $T_D \times N$ matrix. **Intuitively, these are the moments of a regression of Y_D on X_D.**
- For models imposing priors based on **known** and **linear** restrictions, one can find such matrices.
- The posterior moments are:

$$\begin{aligned} \bar{\Omega}^{-1} &= \Omega_0^{-1} + Z'Z = X_D' X_D + Z'Z \\ \bar{B} &= \bar{\Omega}(\Omega_0^{-1} B_0 + X'Y) = (X_D' X_D + X'X)^{-1}(X_D' Y_D + X'Y) \end{aligned}$$

Dummy variable implementation of priors

- **Method 3 (random variance, nonlinear)** Del Negro and Schorfheide. Close to Method 2, but can handle cases where the restrictions towards which the prior shrink can not be easily written using Y_D and X_D, e.g. because they are nonlinear (and possibly depend hyerarchically on some hyperparameter).
- Just add pseudo-observations to the model:

$$\underset{T^*\times N}{\begin{bmatrix} Y \\ Y^* \end{bmatrix}} = \underset{T^*\times k}{\begin{bmatrix} X \\ X^* \end{bmatrix}} \underset{k\times N}{B} + \underset{T^*\times N}{\begin{bmatrix} V \\ V^* \end{bmatrix}}; \; V \sim N\left(0, \begin{matrix} \underset{N\times N}{\Sigma} & 0 \\ 0 & \Sigma^* \end{matrix}\right)$$

- The posterior mean will be:

$$\begin{aligned} & \left(\underset{T^*\times k}{\begin{bmatrix} X' & X'^* \end{bmatrix}} \underset{T^*\times k}{\begin{bmatrix} X \\ X^* \end{bmatrix}}\right)^{-1} \underset{T^*\times k}{\begin{bmatrix} X' & X'^* \end{bmatrix}} \underset{T^*\times N}{\begin{bmatrix} Y \\ Y^* \end{bmatrix}} \\ = \; & (X'X + X^{*\prime}X^*)^{-1}(X'Y + X^{*\prime}Y^*) \end{aligned}$$

Dummy variable implementation of priors

- To remove the stochastic variation, use the population moments.

$$\begin{aligned} B_0 &= (X'X + E[X^{*\prime}X^*])^{-1}(X'Y + E[X^{*\prime}Y^*]) \\ &= (X'X + \Gamma_{[X^{*\prime}X^*]}(\theta))^{-1}(X'Y + \Gamma_{[X^{*\prime}Y^*]}(\theta)) \end{aligned}$$

- This means you **do not** actually need to simulate the artificial data!
- Hyerarchical approach: model the hyperparameters θ

Problems of the conjugate prior

- Conjugate prior is restrictive, as highlighted by Rothenberg (1963), Zellner (1973), Kadiyala and Karlsson (1993, 1997), and Sims and Zha (1998)
- There are many situations in which the form $\Sigma \otimes \Psi_0$ can turn out to be particularly unappealing
- **First**, it prevents permitting any asymmetry in the prior across equations, because the coefficients of each equation feature the same prior variance matrix Ψ_0 (up to a scale factor given by the elements of Σ).
- For example, the traditional Minnesota prior in the original Litterman (1986) implementation can not be cast in such a convenient form, because it imposes cross-variable shrinkage on lags of variables
- Consider the case of a bivariate VAR in the variables y_1 and y_2 and suppose that the researcher has a strong prior belief that y_2 does not Granger cause y_1, while he has not strong beliefs that y_2 itself follows a univariate stationary process. This system of beliefs would require shrinking strongly towards zero the coefficients attached to y_2 in the equation for y_1 and not viceversa.

Problems of the conjugate prior

- **Second**, the Kronecker structure $\Sigma \otimes \Psi_0$ implies the unappealing consequence that prior beliefs must be correlated across the equations of the reduced form representation of the VAR, with a correlation structure proportional to that of the disturbances (as described by the matrix Σ).
- Sims and Zha (1998) discuss in depth this issue, and propose an approach which allows for a more reasonable structure of the coefficient prior variance, which attains computational gains of order $O(N^2)$. Their approach is based on eliciting a prior featuring independence among the *structural* equations of the system, but does not achieve computational gains for an asymmetric prior on the *reduced form* equations coefficients.

Sims and Zha approach

- In particular, the approach of Sims and Zha (1998) achieves conceptual and computational gains by
  - (i) working on the *structural* representation of the VAR, in which the matrix of the errors is diagonal
  - (ii) allowing independence across the coefficients belonging to different *structural* equations, which amounts to the prior variance of the coefficients being block-diagonal, which is desirable as it breaks the unreasonable symmetry across equations implied by the conjugate N-IW prior.
- These two ingredients ensure that the posterior variance matrix has a block-diagonal structure, and therefore achieves computational gains of order N^2.
- However, such strategy still implies that the beliefs about the *reduced form* coefficients are correlated across equations in a way that depends on the covariance of the reduced form errors of the model, and gains are not attainable if one wants to impose an asymmetric prior on these *reduced form* coefficients

The Jeffrey's prior

- This is the limiting case of the conjugate.
- It is specified as:

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

- It delivers:

$$\beta|y, \Sigma \sim N(\beta_1, \Sigma \otimes \Psi_1);\ \Sigma \sim IW(S_1, \nu_1)$$

with:

$$\begin{aligned} \Psi_1 &= (X'X)^{-1};\ \nu_1 = T \\ B_1 &= (X'X)^{-1}(X'Y) = \hat{B}_{OLS} \\ S_1 &= Y'Y - B_1'\Psi_1^{-1}B_1 = Y'Y - Y'X(X'X)^{-1}X'Y = \hat{E}'\hat{E} \end{aligned}$$

where $\hat{E}$ is the matrix of OLS residuals.

The independent N-IW prior

- In this case:

$$\beta \sim N(\beta_0, \Omega_0); \quad \Sigma \sim IW(v_0, S_0)$$

- The conditional posterior of β is

$$\beta | y, \Sigma \sim N(\beta_1, \Omega_1) \tag{7}$$

with

$$\Omega_1 = (\Omega_0^{-1} + (\Sigma^{-1} \otimes X'X))^{-1}$$

$$\beta_1 = \Omega_1 \left(\Omega_0^{-1} \beta_0 + (\Sigma^{-1} \otimes X')y \right)$$

The independent N-IW prior

- The conditional posterior of Σ is

$$\Sigma|y, \beta \sim IW(S_1 = S_0 + S, \nu_1 = \nu_0 + T)$$

where

$$S = (y - (I_N \otimes X)\beta)'(y - (I_N \otimes X)\beta)$$

- The joint posterior $p(\beta, \Sigma|y)$ and the marginals can be obtained by drawing in turn from the conditionals using Gibbs sampling.
- There is no closed form solution for the marginal likelihood.

The independent N-IW prior

- Computational time can be taxing: Consider drawing $m = 1, ..., M$ draws from the posterior of β. To perform a draw β^m from (7), one needs to draw a $N(Np+1)-$dimensional random vector (distributed as a standard Gaussian), denoted rand, and to compute:

$$\beta_1^m = \Omega_1 \left(\Omega_0^{-1} \beta_0 + (\Sigma^{-1} \otimes X') y \right) + chol(\Omega_1) \times \mathrm{rand}, \tag{8}$$

  where $X_t = [1, y'_{t-1}, ..., y'_{t-p}]'$ is the $(Np+1)$-dimensional vector collecting the regressors
- The calculation above involves computations of the order of $4O(N^6)$.

The independent N-IW prior

- Compute:

$$\beta^m = \Omega_1 \left(\Omega_0^{-1} \beta_0 + (\Sigma^{-1} \otimes X') y \right) + chol(\Omega_1) \times \text{rand}, \tag{9}$$

- Indeed, it is necessary to compute:
  - i) the matrix Ω_1 by inverting

$$\Omega_1^{-1} = \Omega_0^{-1} + (\Sigma^{-1} \otimes X'X); \tag{10}$$

  - ii) its Cholesky factor $chol(\Omega_1)$;
  - iii) multiply the matrices obtained in i) and ii) by the vector in the curly brackets of (8) and the vector rand respectively.
- Since each of these operations requires $O(N^6)$ elementary operations, the total computational complexity to compute a draw Π^m is $4 \times O(N^6)$.
- Also computation of $\underline{\Omega}_\Pi^{-1} \text{vec}(\underline{\mu}_\Pi)$ requires $O(N^6)$ operations but this is fixed across repetitions so it needs to be computed just once.

The independent N-IW prior

Some speed improvements can be obtained as follows.

- Define $\Omega_1^{-1} = C'C$ where C is an upper triangular matrix and C' is therefore the Cholesky factor of Ω_1^{-1}. It follows that $\Omega_1 = C^{-1}C'^{-1}$ with C^{-1} upper triangular.
- Clearly, draws from $C^{-1} \times \text{rand}$ will have variance Ω_1 so we can use $C^{-1} \times \text{rand}$ rather than $chol(\Omega_1) \times \text{rand}$.
- Moreover we can substitute $\Omega_1 = C^{-1}C'^{-1}$ in (8) and take C^{-1} as common factor to obtain:

$$\beta^m = C^{-1}\left[C^{-1\prime}\left\{\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes X')y\right\} + \text{rand}\right]. \qquad (11)$$

The independent N-IW prior

- In the above expression C is triangular so its inversion is less expensive, in particular one can simply use the command for backward solution of a linear system as suggested by Chan (2015) instead of inverting the matrices:

$$\beta^m = C \backslash \left[C' \backslash \left\{ \Omega_0^{-1} \beta_0 + (\Sigma^{-1} \otimes X') y \right\} + \text{rand} \right], \tag{12}$$

  where $X = C \backslash B$ is the matrix division of C into B, which is roughly the same as $C^{-1}B$, except it is computed as the solution of the equation $CX = B$.
- A draw in this case still requires the computation of the Cholesky factor of $\bar{\Omega}_{\Pi}^{-1}$ and its inversion, but the multiplications are avoided. Using (12) to perform a draw requires only $2O(N^6)$.
- While this is twice as fast as using (8), it is just a linear improvement and it is not sufficient to solve the bottleneck in estimation of large systems

Triangularization

- Carriero, Clark, Marcellino (2016) introduce an estimation method that solves the problems we discussed above.
- It does so simply by blocking the conditional posterior distribution in N different blocks.
- Recall that in the step of the Gibbs sampler that involves drawing β^m, all of the remaining model coefficients are given, and consider the decomposition $\Sigma = A^{-1}\Lambda A^{-1\prime}$, which gives:

$$v_t = A^{-1}\Lambda^{0.5}\epsilon_t$$

$$\begin{bmatrix} v_{1,t} \\ v_{2,t} \\ \dots \\ v_{N,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ a^*_{2,1} & 1 & & \dots \\ \dots & & 1 & 0 \\ a^*_{N,1} & \dots & a^*_{N,N-1} & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^{0.5} & 0 & \dots & 0 \\ 0 & \lambda_2^{0.5} & & \dots \\ \dots & & \dots & 0 \\ 0 & \dots & 0 & \lambda_N^{0.5} \end{bmatrix} \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \dots \\ \epsilon_{N,t} \end{bmatrix}$$

where $a^*_{j,i}$ and $\lambda_i^{0.5}$ are available under knowledge of Σ.

Triangularization

- We will also denote by $\beta^{(i)}$ the vector of coefficients for equation i contained in row i of B, for the intercept and coefficients on lagged y_t. The VAR can be written as:

$$
\begin{aligned}
y_{1,t} &= \beta_1^{(0)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \beta_{1,l}^{(i)} y_{i,t-l} + \lambda_{1,t}^{0.5}\epsilon_{1,t} \\
y_{2,t} &= \beta_2^{(0)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \beta_{2,l}^{(i)} y_{i,t-l} + a_{2,1}^{*}\lambda_{1,t}^{0.5}\epsilon_{1,t} + \lambda_{2,t}^{0.5}\epsilon_{2,t} \\
&\dots \\
y_{N,t} &= \beta_N^{(0)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \beta_{N,l}^{(i)} y_{i,t-l} + a_{N,1}^{*}\lambda_{1,t}^{0.5}\epsilon_{1,t} + \cdots \\
&\quad + a_{N,N-1}^{*}\lambda_{N-1,t}^{0.5}\epsilon_{N-1,t} + \lambda_{N,t}^{0.5}\epsilon_{N,t},
\end{aligned}
$$

Consider estimating these equations in order from $j = 1$ to $j = N$. When estimating the generic equation j the term on the left hand side is known

Triangularization

- Therefore, we can define:

$$y^*_{j,t} = y_{j,t} - (a^*_{j,1}\lambda^{0.5}_{1,t}\epsilon_{1,t} + ... + a^*_{j,,j-1}\lambda^{0.5}_{j-1,t}\epsilon_{j-1,t}), \tag{13}$$

- The generic equation for variable j is:

$$y^*_{j,t} = \beta^{(0)}_j + \sum_{i=1}^{N}\sum_{l=1}^{p}\beta^{(i)}_{j,l}y_{i,t-l} + \lambda_{j,t}\epsilon_{j,t}. \tag{14}$$

and equation (14) becomes a standard generalized linear regression model for the variables in equation (13), with independent Gaussian disturbances with mean 0 and variance $\lambda_{j,t}$.

Posterior triangularization

- The distribution (7) can be factorized as:

$$\begin{aligned} p(B|A,\Lambda_T,y) &= p(\beta^{(N)}|\beta^{(N-1)},\beta^{(N-2)},\ldots,\beta^{(1)},A,\Lambda_T,y) \\ &\quad \times p(\beta^{(N-1)}|\beta^{(N-2)},\ldots,\beta^{(1)},A,\Lambda_T,y) \\ &\quad \vdots \\ &\quad \times p(\beta^{(1)}|A,\Lambda,y), \end{aligned} \tag{15}$$

  with generic element:

$$\begin{aligned} & p(\beta^{(j)}|\beta^{(j-1)},\beta^{(j-2)},\ldots,\beta^{(1)},A,\Lambda,y) \\ = \;& p(B^{\{j\}}|B^{\{1:j-1\}},A,\Phi,\Lambda,y) \\ \propto \;& p(y|B^{\{j\}},B^{\{1:j-1\}},A,\Lambda)p(B^{\{j\}}|B^{\{1:j-1\}}), \end{aligned}$$

  where $B^{\{j\}} = \beta^{(j)'}$ denotes the (transposed of the) j-th row of the matrix B, and $B^{\{1:j-1\}}$ all of the previous $1,...,j-1$ rows (transposed).
- The term $p(y|B^{\{j\}},B^{\{1:j-1\}},A,\Lambda)$ is the likelihood of equation j
- The term $p(B^{\{j\}}|B^{\{1:j-1\}})$ is the prior on the coefficients of the j-th equation, conditionally on the previous equations.

Posterior moments

- It follows that using the factorization in (15) together with the model in (14) allows one to draw the coefficients of the matrix B in separate blocks $B^{\{j\}}$ which can be obtained from:

$$B^{\{j\}} | B^{\{1:j-1\}}, A, \Lambda, y \sim N(\bar{\mu}_{B^{\{j\}}}, \overline{\Omega}_{B^{\{j\}}}) \tag{16}$$

with

$$\bar{\mu}_{B^{\{j\}}} = \overline{\Omega}_{B^{\{j\}}} \left\{ \underline{\Omega}^{-1}_{B^{\{j\}}} \underline{\mu}_{B^{\{j\}}} + \sum_{t=1}^{T} X_{j,t} \lambda^{-1}_{j,t} y^{*\prime}_{j,t} \right\} \tag{17}$$

$$\overline{\Omega}^{-1}_{B^{\{j\}}} = \underline{\Omega}^{-1}_{B^{\{j\}}} + \sum_{t=1}^{T} X_{j,t} \lambda^{-1}_{j,t} X'_{j,t}, \tag{18}$$

where $y^{*}_{j,t}$ is defined in (13) and where $\underline{\Omega}^{-1}_{B^{\{j\}}}$ and $\underline{\mu}_{B^{\{j\}}}$ denote the prior moments on the j-th equation, given by the j-th column of $\underline{\mu}_{B}$ and the j-th block on the diagonal of $\overline{\Omega}^{-1}_{B}$.

- Note we have implicitly assumed here that the matrix $\underline{\Omega}^{-1}_{B}$ is block diagonal. This assumption can be easily relaxed

MC sampling of B

- The joint posterior distribution of B can be simulated recursively in separate blocks $B^{\{1\}}, B^{\{2\}}|B^{\{1\}}, B^{\{3\}}|B^{\{1:2\}}, ..., B^{\{N\}}|B^{\{1:N-1\}}$ using (16).
- This amounts to simple Monte Carlo simulation
- This MC will produce draws **numerically identical** to those that would be obtained using system-wide estimation
- Any difference in the simulated posterior draws will be due to random variation (which eventually vanishes) and rounding numerical errors.
- The total computational complexity of this estimation algorithm is $O(N^4)$. This is considerably smaller than the complexity of $O(N^6)$ implied by the standard estimation algorithm, with a gain of N^2.

Prior dependence

- There might be cases in which a researcher wishes to specify priors which feature correlations across coefficients belonging to different equations.
- For this case, the general form of the posterior can be obtained easily using a similar triangularization argument on the joint prior distribution, and equation (16) generalizes to:

$$B^{\{j\}}|B^{\{1:j-1\}}, A, \Lambda, y \sim N(\bar{\mu}_{B^{\{j|1:j-1\}}}, \overline{\Omega}_{B^{\{j|1:j-1\}}})$$

with

$$\begin{aligned}
\bar{\mu}_{B^{\{j|1:j-1\}}} &= \overline{\Omega}_{B^{\{j|1:j-1\}}} \left\{ \sum_{t=1}^{T} X_{j,t} \lambda_{j,t}^{-1} y_{j,t}^{*\prime} + \underline{\Omega}^{-1}_{B^{\{j|1:j-1\}}} \underline{\mu}_{B^{\{j|1:j-1\}}} \right\} \\
\overline{\Omega}^{-1}_{B^{\{j|1:j-1\}}} &= \underline{\Omega}^{-1}_{B^{\{j|1:j-1\}}} + \sum_{t=1}^{T} X_{j,t} \lambda_{j,t}^{-1} X_{j,t}^{\prime},
\end{aligned}$$

where $\underline{\mu}_{B^{\{j|1:j-1\}}}$ and $\underline{\Omega}_{B^{\{j|1:j-1\}}}$ are the moments of $B^{\{j\}}|B^{\{1:j-1\}} \sim N(\underline{\mu}_{B^{\{j|1:j-1\}}}, \underline{\Omega}_{B^{\{j|1:j-1\}}})$, i.e. the conditional priors (for equation j conditional on all of the previous equations) implied by the joint prior specification.

Prior dependence

- The conditional prior moments can be obtained recursively using (**??**) and standard results on multivariate Gaussian distributions:

$$\begin{aligned}
\underline{\mu}_{B\{j|1:j-1\}} &= \underline{\mu}_{B\{j\}} + \underline{\Omega}_{B\{[j][1:j-1]\}} \underline{\Omega}^{-1}_{B\{[1:j-1][1:j-1]\}} (B^{\{1:j-1\}} - \underline{\mu}_{B\{1:j-1\}}), \\
\underline{\Omega}_{B\{j|1:j-1\}} &= \underline{\Omega}_{B\{j\}} - \underline{\Omega}_{B\{[j][1:j-1]\}} \underline{\Omega}^{-1}_{B\{[1:j-1][1:j-1]\}} \underline{\Omega}'_{B\{[j][1:j-1]\}}
\end{aligned}$$

  where $\underline{\Omega}_{B\{j\}}$ denotes the block of $\underline{\Omega}_B$ corresponding to equation j, $\underline{\Omega}_{B\{[1:j-1][1:j-1]\}}$ denotes all the blocks on the main block-diagonal, north-west of $\underline{\Omega}_{B\{j\}}$, and $\underline{\Omega}_{B\{[j][1:j-1]\}}$ denotes all the blocks to the left of $\underline{\Omega}_{B\{j\}}$.
- The computational cost of deriving these conditional prior moments is negligible as they need to be computed only once outside the main MCMC sampler.
- Clearly in case of a prior independent across equations $\underline{\Omega}_{B\{[j][1:j-1]\}}$ is a zero matrix and these expressions simplify to $\underline{\mu}_{B\{j|1:j-1\}} = \underline{\mu}_{B\{j\}}$ and $\underline{\Omega}_{B\{j|1:j-1\}} = \underline{\Omega}_{B\{j\}}$, yielding (17) and (18).

The Normal-diffuse (Zellner's) prior

- In this case:

$$\beta \sim N(\beta_0, \Omega_0); \quad p(\Sigma) \propto |\Sigma|^{-\frac{N+1}{2}}$$

- The posteriors are:

$$\begin{aligned} \beta|y, \Sigma &\sim N(\beta_1, \Omega_1) \\ \Sigma|y, \beta &\sim IW(S, T) \end{aligned}$$

where β_1, Ω_1 and S are:

$$\Omega_1 = (\Omega_0^{-1} + (\Sigma^{-1} \otimes X'X))^{-1}$$

$$\beta_1 = \Omega_1 \left(\Omega_0^{-1} \beta_0 + (\Sigma^{-1} \otimes X')y \right)$$

$$S = (y - (I_N \otimes X)\beta)'(y - (I_N \otimes X)\beta)$$