BANK OF ENGLAND

Economic modelling and forecasting

9-13 April 2015

Introduction to multivariate Bayesian estimation

Ole Rummel

Adviser, CCBS at the Bank of England

ole.rummel@bankofengland.co.uk

BANK OF ENGLAND

Outline

- Issues with the basic VAR model: the curse of dimensionality
- Approaches to BVAR estimation:
  - the Minnesota prior; and
  - the Normal-inverse Wishart (natural conjugate) prior
- The Minnesota prior and hyperparameters
- Example of the Minnesota prior
- The likelihood function of a VAR
- The natural conjugate prior
- VARs and Gibbs sampling
- Large and very large BVARs
- Summary

The basic vector autoregression model

- A vector autoregression (VAR) consists of $t = 1, \ldots, T$ observations on a set of n endogenous macroeconomic variables $y_t = (y_{1t}, \ldots, y_{nt})'$, such that y_t is a $(n \times 1)$ vector containing T observations on n time series
- The VAR(p) process with p lags is then defined as:

$$y_t = c + B_1 y_{t-1} + \ldots + B_p y_{t-p} + \varepsilon_t \qquad (1)$$

where:

  - the B_i are $(n \times n)$ coefficient matrices for $i = 1, \ldots, p$;
  - c is a $(n \times 1)$ vector of intercepts; and
  - $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})'$ is an unobservable, $(n \times 1)$ error vector with $E(\varepsilon_t) = 0$ and time-invariant (i.e., constant), positive-definite variance-covariance matrix $cov(\varepsilon_t) = E(\varepsilon_t \varepsilon_t') = \Sigma$ (white noise), such that $\varepsilon_t \sim N(0, \Sigma)$

The basic VAR model: a two-variable VAR(2)

- For example, with $y_t = (y_{1t}, y_{2t})'$, the VAR(2) of the (2×1) vector process for y_t in (1) is:

$$y_{1t} = c_1 + \beta_{11,1} y_{1,t-1} + \beta_{12,1} y_{2,t-1} + \beta_{11,2} y_{1,t-2} + \beta_{12,2} y_{2,t-2} + \varepsilon_{1t}$$

$$y_{2t} = c_2 + \beta_{21,1} y_{1,t-1} + \beta_{22,1} y_{2,t-1} + \beta_{21,2} y_{1,t-2} + \beta_{22,2} y_{2,t-2} + \varepsilon_{2t}$$

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \beta_{11,1} & \beta_{12,1} \\ \beta_{21,1} & \beta_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \beta_{11,2} & \beta_{12,2} \\ \beta_{21,2} & \beta_{22,2} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

$$y_t = c + B_1 y_{t-1} + B_2 y_{t-2} + \varepsilon_t$$

$$\operatorname{cov}(\varepsilon_t) = E(\varepsilon_t \varepsilon_t') = E\left[\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \begin{pmatrix} \varepsilon_{1t} & \varepsilon_{2t} \end{pmatrix} \right] = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{bmatrix} = \Sigma$$

The basic VAR model: notation and definitions (1)

- There are many alternative ways of writing the VAR(*p*) given by (1)
- For example, using the format of a simultaneous equations system, the basic VAR model can be re-written in matrix form in two different ways (Canova (2007), Koop and Korobilis (2009)):
  - one approach expresses the model in terms of the multivariate Normal distribution; while
  - the other expresses the model in terms of the matric-variate Normal distribution

The basic VAR model: notation and definitions (1)

- Common to both approach are the following definitions:

$$x_t = \begin{pmatrix} 1 \\ y'_{t-1} \\ \vdots \\ y'_{t-p} \end{pmatrix} \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix} \quad B = \begin{pmatrix} c \\ B_1 \\ \vdots \\ B_p \end{pmatrix}$$

  and we also need $y = \text{vec}(y_t)$, $b = vec(B)$ and $\varepsilon = \text{vec}(\varepsilon_t)$

- If $k = 1 + np$ is the number of coefficients in each equation of the VAR, X will be a $(T \times k)$ matrix

The basic VAR model: notation and definitions (2)

- The multivariate Normal distribution approach arises if we use an ($nT \times 1$) vector $y = (y_1', y_2', \ldots, y_T')$ which stacks all T observations on the first dependent variable, all T observations on the second dependent variable, etc.:

$$y_{(nT \times 1)} = (\mathrm{I}_n \otimes X)\, b + \varepsilon \qquad \varepsilon \sim \mathrm{N}(0, \Sigma \otimes \mathrm{I}_n) \tag{2}$$

  and ε is stacked conformably
- As we will see later, (2) is useful for decomposing the likelihood function of a VAR(p) into the product of a Normal density for b, conditional on the OLS estimates of the VAR parameters (b_{OLS}) and Σ, and an (inverse) Wishart density for Σ

The basic VAR model: notation and definitions (2)

- The matric-variate Normal distribution approach arises if we define Y to be a $(T \times n)$ matrix which stacks the T observations on each dependent variable in columns next to one another:

$$Y_{(T \times n)} = X_{(T \times (1+np))} B_{((1+np) \times n)} + E_{(T \times n)} \qquad E \sim \text{MN}(0, \Sigma \otimes \text{I}_n) \qquad (3)$$

- While b is a $(kn \times 1)$ vector of VAR coefficients in the VAR given by (2), now B is a $(k \times n)$ matrix of VAR coefficients
- The relationship between the two is that $b = \text{vec}(B)$

The basic VAR model: estimation

- In the absence of any restrictions, each equation in a VAR has the same regressors, and each equation in our VAR(2) example may be estimated **separately** by ordinary least squares (OLS):

$$y_t = c + B_1 y_{t-1} + B_2 y_{t-2} + \varepsilon_t$$

$$\hat{\beta} \equiv \begin{bmatrix} c \\ B_1 \\ B_2 \end{bmatrix} = (X'X)^{-1} X'Y \qquad (4)$$

where $Y = \begin{bmatrix} y_{1,3} & y_{2,3} \\ y_{1,4} & y_{2,3} \\ \vdots & \vdots \\ y_{1,T} & y_{2,T} \end{bmatrix}$ *and* $X = \begin{bmatrix} 1 & y_{1,2} & y_{2,2} & y_{1,1} & y_{2,1} \\ 1 & y_{1,3} & y_{2,3} & y_{1,2} & y_{2,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & y_{1,T-1} & y_{2,T-1} & y_{1,T-2} & y_{2,T-2} \end{bmatrix}$

Issues with the basic VAR(*p*) model

- The number of coefficients in the basic (reduced-form) VAR(*p*) model in (1) easily proliferates, meaning that:
  - there are lots of coefficients to estimate and so these models are over-parameterised – the total number of parameters to be estimated equals $n(np + 1)$;
  - unrestricted OLS estimates of the coefficient values are often not very well determined (in the sense of imprecise) in a finite set of data; and
  - if the $B_1, \ldots, B_p$ are imprecisely estimated because of limited data, the impulse response functions, forecasts and forecast error variance decompositions that are based on them will also be imprecisely estimated
- Alternative methods for estimating the coefficients have therefore been developed (…with a common theme being '**shrinkage**')

Dealing with the curse of dimensionality

- To reduce the number of parameters in a high-dimensional VAR, one could of course set ('shrink') many coefficients equal to a particular values, such as zero, or impose the condition that the same coefficient interacts with multiple regressors
- This can be done either in the form of **hard** or **soft** restrictions:
  - unfortunately, **hard** (zero) restrictions might rule out the existence of certain spillover effects, which could be undesirable; while
  - the use of **soft** restrictions, on the other hand, is conceptually more appealing – these can be easily incorporated through assigning probability distributions on coefficients that are 'centred' on the desired restrictions, but that have a small, yet non-zero, variance
- The latter approach involves the application of **Bayesian** techniques

Fundamentals of Bayesian econometrics (1)

- Classical econometrics treats the parameters of a model as **fixed**, unknown constants to be estimated
- Within the Bayesian framework, the parameters of the model are treated as **random variables** that have probability distributions
- These distributions are used to summarise the status of knowledge about the parameters of the model…this is frequently possible even **before** the estimation procedure
- Bayesian estimation methods (and the use of prior information) provide a logical and formally consistent way of introducing shrinkage

Fundamentals of Bayesian econometrics (2)

- In Bayesian inference, a **prior distribution** is updated by sample information contained in the likelihood function to form a **posterior distribution**
- Thus, to the extent that the prior is based on **non-sample** information, it provides the ideal framework for containing different sources of information and thereby sharpening inference on macroeconometric analysis

Bayes' theorem

- Bayes' theorem allows us to compute the posterior distribution of a generic parameter vector, θ, $p(\theta|y)$, from the prior distribution, $p(\theta)$, and the likelihood function, $L(y|\theta)$, where the **posterior** distribution is proportional to the **likelihood function** times the **prior** distribution:

$$p(\theta|y) \propto L(y|\theta) \times p(\theta) \quad (5)$$

posterior information	$\propto$	sample information	$\times$	prior information

- The classical approach is based on sample information only (i.e., the likelihood function, $L(y|\theta)$)
- In contrast, the Bayesian approach **combines** the sample information in $L(y|\theta)$ with the researcher's beliefs (embodied in $p(\theta)$)

Approaches to BVAR estimation

- The Bayesian approach suggest a solution to the curse of dimensionality by introducing **prior distributions**
- A variety of priors can be used within the VAR approach – they differ mainly in whether they lead to analytical results for the posterior and predictive densities or whether Markov-chain Monte Carlo methods (such as Gibbs sampling and the Metropolis-Hastings algorithm) are required to carry out Bayesian estimation

Approaches to BVAR estimation

- There are different approaches to estimating Bayesian VARs (BVARs):
  - Litterman's (1980, 1986) 'Minnesota' (reference) prior;
  - the Normal-inverse Wishart (natural conjugate) prior;
  - the independent Normal-inverse Wishart prior;
  - the diffuse prior;
  - the steady-state prior; etc.

A Bayesian reference prior (1)

- A **reference** prior must consider several partially conflicting aspects of actual econometric practice:
  - the number of parameters in VAR models is usually very large and it is not realistic to demand a detailed subjective specification of prior distributions on such high-dimensional spaces – a prior with relatively few **hyperparameters**, each with a clear interpretation, is thus called for;
  - priors that are not transparent, in the sense that one cannot easily understand the kind of information they convey, should not be used in practice;

A Bayesian reference prior (2)

- A **reference** prior must consider several partially conflicting aspects of actual econometric practice:
  - the prior must lead to straightforward posterior calculations which can be performed on a routine basis without the need for fine-tuning in each new application; and
  - a prior with the above requirements will probably not coincide with the investigator's actual prior beliefs, but should nevertheless be useful as a point of reference or an agreed standard

Litterman's Bayesian prior

- Litterman (1980,1986) specifies his 'Minnesota' prior by appealing to three statistical regularities of macroeconomic time series data:
  - the typical trending behaviour of macroeconomic time series, i.e., the fact that many economic time series contain a stochastic trend (unit root);
  - the fact that more recent values of a series usually contain more information about the current value of the series than past values; and
  - the fact that past values of the variable itself contain more information about its current value of the series than past values of other variables
- These statistical regularities allow us to formulate hypotheses regarding the **means** of the prior distributions for the coefficients

The Minnesota prior: the prior mean (1)

- Many macroeconomic time series follow a random walk, meaning that:
  - the prior mean of the coefficient on the own first lag of each variable is one; and
  - the prior means of coefficients on higher lags (cross lags) are likely to be close to zero
- Obviously, this prior mean does not make sense for all types of series
- Note that a random walk prior mean may be inappropriate for a stationary series – in practice, this choice should depend on the stationarity properties of the underlying time series, i.e., we can set prior means less than one in magnitude (or even zero for stationary series)

The Minnesota prior: the prior mean (2)

- If the assumptions of the Minnesota prior were (strictly) true, the VAR(2) of the (2×1) vector process for y_t in (1) would be given by:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \quad (6)$$

  such that both y_{1t} and y_{2t} are random walks (in this case without a drift – constant – term):

$$y_{1t} = y_{1,t\text{-}1} + \varepsilon_{1t}$$
$$y_{2t} = y_{2,t\text{-}1} + \varepsilon_{2t}$$

- While the random walk prior in (6) might be considered a reasonable specification, there is no need to impose it exactly on the VAR

The Minnesota prior: the prior variance

- According to the Minnesota prior, the prior variance for coefficient β_{ij} is specified as follows:

$$\left(\frac{\lambda_1}{p^{\lambda_3}}\right)^2 \quad \text{if } i = j \qquad (7)$$

$$\left(\frac{\sigma_i \lambda_1 \lambda_2}{\sigma_j p^{\lambda_3}}\right)^2 \quad \text{if } i \neq j \qquad (8)$$

$$(\sigma_i \lambda_4)^2 \quad \text{for the constant} \qquad (9)$$

  where p is the lag length; λ_1, λ_2 , λ_3 and λ_4 are **hyperparameters** set by the researcher; and σ_i and σ_j are the variances of the error term in equations i and j – i refers to the dependent variable in the i-th equation and j to the independent variables in that equation

The Minnesota prior: the hyperparameters (1)

- Technically, the requirements of the Minnesota prior are expressed formally by introducing a vector of **hyperparameters** $\Lambda = (\lambda_1, \ldots, \lambda_r)$…
- …which apply to **all** coefficients to be estimated
- The **prior variance-covariance matrix** for the VAR coefficients in each equation is then specified in terms of these hyperparameters
- In essence, this cuts down the number of estimated parameters in the VAR from $n(np + 1)$ to the number of hyperparameters, r
- Note that this will essentially be true for a VAR of any dimension n and p

The Minnesota prior: the hyperparameters (2)

- For instance, the Minnesota prior has four hyperparameters:
  - the hyperparameter λ_1 controls the standard deviation of the prior on its own lags (β_{ii}), and thus how closely the random walk approximation is to be imposed;
  - the hyperparameter λ_2 ($0 < \lambda_2 \leq 1$) controls the standard deviation of the prior on lags of variables other than the dependent variables, i.e., lags of variable i in equation j (β_{ij});
  - the hyperparameter $\lambda_3 > 0$ controls the degree to which coefficients on lags higher than one are likely to be zero;
  - the prior variance on the constant in equation i is defined as $\sigma_i \lambda_4$; and
  - the ratio σ_i/σ_j accounts for different scales of the variables in equation i and equation j (in practice, the σ_i are usually set to the residual standard error from an OLS regression of each dependent variable on p lagged values)

The Minnesota prior: the hyperparameters (3)

- Settings for the hyperparameters in the Minnesota prior that have been found to work well in practice can be found in Doan *et al.* (1984), Robertson and Tallman (1999) and Canova (2007) – although hyperparameters may also be chosen to maximise the VAR's forecast performance (over what horizon?) rather than being assigned arbitrary values
- Canova (2007, p. 380), for example, recommends:

$$\lambda_1 = 0.2$$

$$\lambda_2 = 0.5$$

$$\lambda_3 = 1 \text{ or } 2$$

$$\lambda_4 = 10^5$$

for the Minnesota prior

The Minnesota prior: illustration (1)

- According to the Minnesota prior, the weight on your prior belief about coefficients in, say, the **first** equation of the VAR(2) of the (2×1) vector process for y_{1t} in (1) would be given by:

$$y_{1t} = c_1 + \beta_{11,1} y_{1,t-1} + \beta_{12,1} y_{2,t-1} + \beta_{11,2} y_{1,t-2} + \beta_{12,2} y_{2,t-2} + \varepsilon_{1t}$$

$$\updownarrow$$

$$(\sigma_1 \lambda_4)^2$$

- As λ_4 goes to zero, the constant c_1 is shrunk toward zero (σ_1^2 is the variance of ε_{1t})

The Minnesota prior: illustration (2)

- According to the Minnesota prior, the weight on your prior belief about coefficients in, say, the **first** equation of the VAR(2) of the (2×1) vector process for y_{1t} in (1) would be given by:

$$y_{1t} = c_1 + \beta_{11,1} y_{1,t-1} + \beta_{12,1} y_{2,t-1} + \beta_{11,2} y_{1,t-2} + \beta_{12,2} y_{2,t-2} + \varepsilon_{1t}$$

$$\updownarrow$$

$$(\lambda_1)^2$$

- As λ_1 goes to zero, the diagonal elements of B_1 ($\beta_{11,1}$ in the first equation and $\beta_{22,1}$ in the second equation) are shrunk toward one and all other coefficients are shrunk to zero

The Minnesota prior: illustration (3)

- According to the Minnesota prior, the weight on your prior belief about coefficients in, say, the **first** equation of the VAR(2) of the (2×1) vector process for y_{1t} in (1) would be given by:

$$y_{1t} = c_1 + \beta_{11,1} y_{1,t-1} + \beta_{12,1} y_{2,t-1} + \beta_{11,2} y_{1,t-2} + \beta_{12,2} y_{2,t-2} + \varepsilon_{1t}$$

$$\updownarrow$$

$$(\sigma_1 \lambda_1 \lambda_2 / \sigma_2)^2$$

- Decreasing λ_2 toward zero has the effect of shrinking the off-diagonal elements of B_i (in this case, $\beta_{12,1}$ and $\beta_{12,2}$) toward zero
- Setting $\lambda_2 = 1$ means that no distinction is made between the lags of the dependent variable and the lags of other variables

The Minnesota prior: illustration (4)

- According to the Minnesota prior, the weight on your prior belief about coefficients in, say, the **first** equation of the VAR(2) of the (2×1) vector process for y_{1t} in (1) would be given by:

$$y_{1t} = c_1 + \beta_{11,1} y_{1,t-1} + \beta_{12,1} y_{2,t-1} + \beta_{11,2} y_{1,t-2} + \beta_{12,2} y_{2,t-2} + \varepsilon_{1t}$$

$$(\lambda_1 / 2^{\lambda_3})^2 \qquad (\sigma_1 \lambda_1 \lambda_2 / \sigma_2 2^{\lambda_3})^2$$

- As λ_3 gets bigger, coefficients on higher lags, such as $\beta_{11,2}$ and $\beta_{12,2}$, are shrunk towards zero more tightly, i.e., they go to zero faster

The Minnesota prior: illustration (5)

- If $n = 2$ and there are $p = 2$ lags of each variable in the VAR, the prior for the coefficients of the **first** equation is:

$$N\left(\begin{pmatrix}0\\1\\0\\0\\0\end{pmatrix},\begin{pmatrix}(\sigma_1\lambda_4)^2 & 0 & 0 & 0 & 0\\ 0 & \lambda_1^2 & 0 & 0 & 0\\ 0 & 0 & (\sigma_1\lambda_1\lambda_2/\sigma_2)^2 & 0 & 0\\ 0 & 0 & 0 & (\lambda_1/2^{\lambda_3})^2 & 0\\ 0 & 0 & 0 & 0 & [\sigma_1\lambda_1\lambda_2/(2^{\lambda_3}\sigma_2)]^2\end{pmatrix}\right)$$

The Minnesota prior: illustration (6)

- If $n = 2$ and there are $p = 2$ lags of each variable in the VAR, the prior for the coefficients of the **second** equation is:

$$N\left(\begin{pmatrix}0\\0\\1\\0\\0\end{pmatrix},\begin{pmatrix}(\sigma_2\lambda_4)^2 & 0 & 0 & 0 & 0\\ 0 & (\sigma_2\lambda_1\lambda_2/\sigma_1)^2 & 0 & 0 & 0\\ 0 & 0 & \lambda_1^2 & 0 & 0\\ 0 & 0 & 0 & [\sigma_2\lambda_1\lambda_2/(2^{\lambda_3}\sigma_1)]^2 & 0\\ 0 & 0 & 0 & 0 & (\lambda_1/2^{\lambda_3})^2\end{pmatrix}\right)$$

The Minnesota prior: set-up (1)

- The prior mean for the (vectorised) VAR coefficients is given by:

$$\tilde{b}_0 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

The Minnesota prior: set-up (2)

- The prior variance matrix for the VAR coefficients is given by:

$$H = \begin{pmatrix}
(\sigma_1\lambda_4)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \left(\frac{\sigma_1\lambda_1\lambda_2}{\sigma_2}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \left(\frac{\sigma_1\lambda_1\lambda_2}{\sigma_2 2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & (\sigma_2\lambda_4)^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2\lambda_1\lambda_2}{\sigma_1}\right)^2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2\lambda_1\lambda_2}{\sigma_1 2^{\lambda_3}}\right)^2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2
\end{pmatrix}$$

The Minnesota prior: estimation (1)

- The Minnesota prior for the VAR coefficients assumes a Normal prior with mean zero and small standard deviations for long lags (reflecting the fact that the larger the lag, the more likely the coefficient is to be close to zero); but allows the data to override this assumption if contrary evidence about a coefficient is strong
- Under this assumption, the (Normal) mean of the posterior distribution for the coefficients is available in closed form (Kadiyala and Karlsson (1997), Table 1)
- Note that the Minnesota prior is informative on **all** the coefficients of the *c* vector as well as the B_i matrices...
- ...but **non-informative** on the other parameters of the model (such as the elements of the variance-covariance matrix, Σ)

The Minnesota prior: estimation (2)

- The Minnesota prior distribution for the VAR coefficients is given by:

$$p(b) \sim N(\tilde{b}_0, H) \tag{10}$$

- A big advantage of the Minnesota prior is that it leads to simple posterior inference involving only a Normal distribution N(M^*, V^*)
- The **mean** of this conditional posterior Normal distribution, which is formally the Bayesian estimate of the VAR coefficients, is:

$$\underset{(n(np+1)\times 1)}{M^*} = (H^{-1} + \hat{\Sigma}^{-1} \otimes X'X)^{-1}(H^{-1}\tilde{b}_0 + \hat{\Sigma}^{-1} \otimes X'X\hat{b}) \tag{11}$$

  where $\hat{\Sigma}$ is the estimated variance-covariance matrix of the VAR
- Note how (11) collapses to the OLS estimate of the VAR coefficients, $\hat{b}$, in the absence of prior information

The Minnesota prior: estimation (3)

- The Minnesota prior distribution for the VAR coefficients is given by:

$$p(b) \sim N(\tilde{b}_0, H)$$

- A big advantage of the Minnesota prior is that it leads to simple posterior inference involving only a Normal distribution N(M^*, V^*)
- The **variance** of this conditional posterior Normal distribution is:

$$\underset{(n(np+1)\times n(np+1))}{V^*} = (H^{-1} + \hat{\Sigma}^{-1} \otimes X'X)^{-1} \quad (12)$$

- Note how (12) collapses to the OLS estimate of the variance of the VAR coefficients in the absence of prior information

The Minnesota prior: estimation (4)

- In light of (11) and (12), the influence of the Minnesota prior distribution on the posterior results of the VAR coefficients can be seen to be twofold:
  - the precision of the 'estimates' is improved because of the usual adding up of prior and sample precision; and
  - the posterior means of the coefficients on which the sample is weakly informative are shrunk towards the prior means (most of them being null), and away from the least-squares estimates (which would be the posterior means under a non-informative prior)
- In VAR models, it is rather customary to find least-squares estimates which are very imprecisely determined, so the prior may help to shrink these coefficients to less extreme values than the least-squares values...
- (...which usually helps to improve the forecasts of the model)

The Minnesota prior: the variance-covariance matrix

- The Minnesota prior is based on an approximation that leads to great simplification in prior elicitation and computation
- The simplicity of the Minnesota prior comes from the fact that Σ is assumed to be known, meaning that **no** prior distribution for Σ is required
- In fact, the simplification involves replacing Σ with an estimate $\hat{\Sigma}$
- Moreover, under a strict interpretation of the Minnesota prior, the variance-covariance matrix of the residuals of the VAR, $\hat{\Sigma}$, is simplified even further by assuming that it is **diagonal** with the diagonal entries fixed using the OLS estimate of the error variance, s_i^2, from each equation (such that $\hat{\sigma}_{ii} = s_i^2$)
- When Σ is replaced by an estimate, we only have to worry about a prior for *b*

The Minnesota prior: the variance-covariance matrix

- One disadvantage of the Minnesota prior is that it does not provide a full Bayesian treatment of Σ as an unknown parameter...
- ...and – given that it simply replaces Σ by $\hat{\Sigma}$, ignoring any uncertainty in this parameter – it will be impossible to adequately represent parameter uncertainty in posterior and predictive densities
- It is therefore common practice amongst some researchers to incorporate the Minnesota prior into a Gibbs sampling algorithm framework and draw Σ from the inverse Wishart distribution

The Minnesota prior: summary

- The two main features of the Minnesota prior, strictly interpreted, are:
  - the posterior independence between equations; and
  - the fixed and diagonal residual variance-covariance matrix (which is, however, often ignored in practice)
- The coefficient problem is therefore simplified because it avoids having to specify how the prior distribution for the variance-covariance matrix, Σ, is related to the prior distribution for the VAR coefficients in B
- These restrictions can obviously be relaxed, giving rise to other priors for VAR estimation that incorporate a full variance-covariance matrix

The likelihood function of the VAR

- The likelihood function of the VAR in (1) is given by:

$$L(b,\Sigma) \propto |\Sigma|^{-T/2}\exp\{-\tfrac{1}{2}\mathrm{tr}[(Y - XB)'\Sigma^{-1}(Y - XB)]\} \qquad (13)$$

- After 'some' manipulation (Kadiyala and Karlsson (1997), Koop *et al.* (2007, p. 314)), this likelihood can be decomposed into the product of a (Normal) distribution for b conditional on $\hat{b}$, the OLS estimate of b, and Σ as well as an (inverse) Wishart distribution for Σ:

$$L(b,\Sigma) \propto N(b \mid \hat{b}, \Sigma \otimes (X'X)^{-1}) \times IW(\Sigma \mid (Y - X\hat{B})'(Y - X\hat{B}), T - k - n - 1) \qquad (14)$$

  where $k = 1 + np$

- As we will see, this decomposition is extremely useful as it suggests (conditional) prior distributions for b and Σ

The natural conjugate prior (1)

- Natural conjugate priors are those where the prior, likelihood and posterior all come from the same family of distributions
- The decomposition of the likelihood suggests that, for the VAR, the natural conjugate priors have the form:

$$p(b \mid \Sigma) \sim N(\tilde{b}_0, \Sigma \otimes \bar{H}) \quad (15)$$

for the VAR coefficients, b, and:

$$p(\Sigma) \sim IW(\bar{S}, \alpha) \quad (16)$$

for the variance-covariance matrix, Σ, where $\bar{S}$ is a prior scale matrix and α are the prior degrees of freedom

- This is a special case of the **Normal-inverse Wishart** prior, which assumes a Normal prior for the VAR coefficients and an inverse Wishart prior for the covariance matrix

The natural conjugate prior (2)

- The prior mean for the (vectorised) VAR coefficients is given by:

$$\tilde{b}_0 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

The natural conjugate prior (3)

- In (15), the matrix $\overline{H}$ is a diagonal matrix, with diagonal elements given by:

$$\left(\frac{\lambda_0\lambda_1}{p^{\lambda_3}\sigma_i}\right)^2 \text{ for the coefficients on lag } p \qquad (17)$$

$$(\lambda_0\lambda_4)^2 \text{ for the constant} \qquad (18)$$

where p is the lag length, λ_0, λ_1, λ_3 and λ_4 are hyperparameters set by the researcher and σ_i is the variance of the error term in equations i , where i refers to the dependent variable in the i-th equation

The natural conjugate prior: the hyperparameters

- The natural conjugate prior also has four hyperparameters:
  - the hyperparameter λ_0, which enters every single prior distribution, controls the overall tightness of the prior on the variance-covariance matrix;
  - the hyperparameter λ_1 controls the standard deviation of the prior on the coefficients on the first lag – as $\lambda_1 \rightarrow 0$, the prior is imposed more tightly;
  - the hyperparameter λ_3 controls the degree to which higher lags ($p > 1$) are likely to be zero - as λ_3 increases, coefficients on higher lags are shrunk to zero more tightly; and
  - the hyperparameter λ_4 controls the prior variance on the constant - as $\lambda_4 \rightarrow 0$, the constant is shrunk to zero

The natural conjugate prior (4)

- So, for the case of our VAR(2) example, the matrix $\overline{H}$ is given by:

$$\overline{H} = \begin{pmatrix} (\lambda_0\lambda_4)^2 & 0 & 0 & 0 & 0 \\ 0 & \left(\frac{\lambda_0\lambda_1}{\sigma_1}\right)^2 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{\lambda_0\lambda_1}{\sigma_2}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & \left(\frac{\lambda_0\lambda_1}{2^{\lambda_3}\sigma_1}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & \left(\frac{\lambda_0\lambda_1}{2^{\lambda_3}\sigma_2}\right)^2 \end{pmatrix} \quad (19)$$

The natural conjugate prior (5)

- The matrix $\overline{S}$ is an ($n \times n$) diagonal matrix, with diagonal elements given by:

$$\left(\frac{\sigma_i}{\lambda_0}\right)^2 \tag{20}$$

such that, for our VAR(2) example, this matrix is given by:

$$\overline{S} = \begin{pmatrix} \left(\frac{\sigma_1}{\lambda_0}\right)^2 & 0 \\ 0 & \left(\frac{\sigma_2}{\lambda_0}\right)^2 \end{pmatrix} \tag{21}$$

The natural conjugate prior (6)

- The prior variance-covariance matrix for the VAR coefficients, *H*, involves calculating $\overline{S} \otimes \overline{H}$:

$$\begin{pmatrix} \left(\frac{\sigma_1}{\lambda_0}\right)^2 & 0 \\ 0 & \left(\frac{\sigma_2}{\lambda_0}\right)^2 \end{pmatrix} \otimes \begin{pmatrix} (\lambda_0\lambda_4)^2 & 0 & 0 & 0 & 0 \\ 0 & \left(\frac{\lambda_0\lambda_1}{\sigma_1}\right)^2 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{\lambda_0\lambda_1}{\sigma_2}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & \left(\frac{\lambda_0\lambda_1}{2^{\lambda_3}\sigma_1}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & \left(\frac{\lambda_0\lambda_1}{2^{\lambda_3}\sigma_2}\right)^2 \end{pmatrix}$$

The natural conjugate prior (7)

- This calculation yields:

$$H = \begin{pmatrix}
(\sigma_1\lambda_4)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \left(\frac{\sigma_1\lambda_1}{\sigma_2}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \left(\frac{\sigma_1\lambda_1}{\sigma_2 2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & (\sigma_2\lambda_4)^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2\lambda_1}{\sigma_1}\right)^2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2\lambda_1}{\sigma_1 2^{\lambda_3}}\right)^2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2
\end{pmatrix}$$

The natural conjugate prior (8)

- Comparing slides (47) and (31), we find that the prior variance-covariance matrix for the natural conjugate prior is the same as that for the Minnesota prior, except for the fact that the hyperparameter $\lambda_2 = 1$ in the former
- What did hyperparameter λ_2 do?
- This parameter controlled the standard deviation of the prior on lags of variables other than the dependent variables, i.e., lags of variable *i* in equation *j* (β_{ij})
- In light of the above, the structure of the natural conjugate prior implies that we treat lags of the dependent variable and lags of the other variables in each equation of the VAR in exactly the same manner – which may be rather **restrictive** in many practical circumstances

The natural conjugate prior (9)

- Given the natural conjugate prior, analytical results exist for the posterior distribution for the coefficients and the covariance matrix
- One clear advantage of this set up over the Minnesota prior is that it allows the derivation of analytical results without the need for a fixed and diagonal error covariance matrix
- The exact formulas for the posteriors are listed in Table 1 in Kadiyala and Karlsson (1997)
- Posterior inference about the VAR coefficients can be carried out by using the fact that the marginal posterior for *b* (i.e., after integrating out Σ) is a multivariate *t*-distribution

The Sims and Zha (1998) prior

- Sims and Zha (1998) provide an example of a natural conjugate prior, define $\underline{H}$ and $\underline{S}$ and show how this prior is related to the Minnesota prior
- Typical hyperparameters for the Sims and Zha (1998) prior are:

Parameter	*Range*	*Interpretation*
λ_0	[0,1]	Overall scale of the error covariance matrix
λ_1	> 0	Standard deviation around A_1 (persistence)
λ_2	$= 1$	Weight of own lag versus other lags
λ_3	> 0	Lag decay
λ_4	≥ 0	Scale of standard deviation of intercept
λ_5	≥ 0	Scale of standard deviation of exogenous variable coefficients
μ_5	≥ 0	Sum of coefficients / Cointegration (long term trends)
μ_6	≥ 0	Initial observations / dummy observation (impacts of initial conditions)
ν	> 0	Prior degrees of freedom

VARs and Gibbs sampling

- In some circumstances – listed in Kadiyala and Karlsson (1997) – exact formulae for posterior moments are available
- Most of the time, though, exact analytical or closed-form solutions do not exist and have to be simulated
- One such simulation technique is the Gibbs sampling algorithm, which is convenient as it provides point estimates **and** measures of uncertainty
- In other words, Bayesian simulation methods such as Gibbs sampling provide an efficient way not only to obtain point estimates but also to characterise the uncertainty around those point estimates

Gibbs sampling (1)

- Gibbs sampling is a simulation method to approximate the posterior distribution of θ without analytical integration
- For the Gibbs sampler to work we need to know the **full conditional posterior distributions**:

$$p(\theta_1|\theta_2, \theta_3, \dots, \theta_B, y)$$
$$p(\theta_2|\theta_1, \theta_3, \dots, \theta_B, y)$$
$$\dots$$
$$p(\theta_B|\theta_1, \theta_2, \dots, \theta_{B\text{-}1}, y)$$

- Consider a problem with two parameters θ_1 and θ_2 and let y denote the data
- Suppose further that we know the conditional distributions $p(\theta_1|\theta_2,y)$ and $p(\theta_2|\theta_1,y)$
- We need to find $p(\theta_1|y)$ and $p(\theta_2|y)$

Gibbs sampling (2)

- Gibbs sampling involves drawing sequentially from the full conditional posterior distributions and works in the following steps
- Choose some initial point $(\theta_1^{(0)}, \theta_2^{(0)})$ from the parameter space – this can be any reasonable value of θ
- Take draws from the two conditional distributions in the following sequence:

$$\theta_1^{(1)} \sim p(\theta_1 | \theta_2^{(0)}, y)$$

$$\theta_2^{(1)} \sim p(\theta_2 | \theta_1^{(1)}, y)$$

- This completes **one** iteration which yields $(\theta_1^{(1)}, \theta_2^{(1)})$

Gibbs sampling (3)

- Next, we use the new parameters $\theta_1^{(1)}$ and $\theta_2^{(1)}$ as starting values and repeat the iteration of random draws:

$$\theta_1^{(2)} \sim p(\theta_1 | \theta_2^{(1)}, y)$$
$$\theta_2^{(2)} \sim p(\theta_2 | \theta_1^{(2)}, y)$$

  to complete another iteration which yields $(\theta_1^{(2)}, \theta_2^{(2)})$
- This sequence of draws is a Markov chain because the values at step *t* only depend on the values at step *t*-1
- If allowed to run long enough (meaning a large enough number of draws), the empirical distributions of the series of draws from the Gibbs sampler will converge to the true posterior distributions, $p(\theta_1|y)$ and $p(\theta_2|y)$

Gibbs sampling (4)

- If allowed to run for sufficiently long after convergence, the Gibbs sampler produces a complete sample from the distribution of θ
- In practice, we run the Gibbs sampler for S replications
- But the first S_0 of these − the so-called **burn-in replications** − are discarded in order to eliminate the effect of the initial values and the remaining S_1 retained for the approximation of $p(\theta|y)$, where $S_0 + S_1 = S$
- The Gibbs sampler is a powerful tool for posterior simulation which is used in many econometric models - but it does require knowledge of the full conditional posterior distributions
- A method that does away with this requirement is the Metropolis-Hastings algorithm – both Gibbs and Metropolis-Hastings are known as Markov-chain Monte Carlo (MCMC) methods

Gibbs sampling algorithm for a VAR (1)

- The Gibbs sampling algorithm for a VAR model consists of the following steps
- **Step 1**: set priors for the VAR coefficients and covariance matrix
- As discussed above, the prior for the VAR coefficients is Normal and is given by

$$p(b) \sim N(\tilde{b}_0, H)$$

- The prior for the variance-covariance matrix of the residuals, Σ, is inverse Wishart and given by $p(\Sigma) \sim IW(\overline{S}, \alpha)$
- Set a starting value for Σ, such as its OLS estimate, for example

Gibbs sampling algorithm for a VAR (2)

- **Step 2**: sample the VAR coefficients from their conditional posterior distribution $p(b|\Sigma, Y_t) \sim \mathrm{N}(M^*, V^*)$, where

$$\underset{(n(np+1))\times 1}{M^*} = (H^{-1} + \Sigma^{-1} \otimes X_t' X_t)^{-1}(H^{-1}\tilde{b}_0 + \Sigma^{-1} \otimes X_t' X_t \hat{b}) \quad (22)$$

$$\underset{(n(np+1))\times(n(np+1))}{V^*} = (H^{-1} + \Sigma^{-1} \otimes X_t' X_t)^{-1} \quad (23)$$

- Once M^* and V^* are calculated, the VAR coefficients are drawn from the Normal distribution:

$$\underset{(n(np+1))\times 1}{b^1} = \underset{(n(np+1))\times 1}{M^*} + \left[\underset{1\times(n(np+1))}{\bar{b}} \times \underset{(n(np+1))\times(n(np+1))}{(V^*)^{1/2}} \right]' \quad (24)$$

where $\bar{b} \sim N(0,1)$, i.e., it is a draw from a standard Normal

Gibbs sampling algorithm for a VAR (3)

- **Step 3**: Draw Σ from its conditional distribution $g(\Sigma \mid b, Y_t) \sim IW(\overline{\Sigma}, T + \alpha)$, where $\overline{\Sigma} = \overline{S} + (Y_t - X_t B^1)'(Y_t - X_t B^1)$
- B^1 is the previous draw of the VAR coefficients, reshaped into a matrix with dimensions $((np + 1) \times n)$ so that it is conformable with X_t
- In terms of the algorithm, to draw a matrix $\hat{\Sigma}$ from the IW distribution with v degrees of freedom and scale parameter S, draw a matrix Z with dimensions $(v \times n)$ from the multivariate Normal $N(0, S^{-1})$
- Then the draw from the inverse Wishart distribution is given by the following transformation:

$$\hat{\Sigma} = \left(\sum_{i=1}^{v} Z_i Z_i' \right)^{-1} \tag{25}$$

Gibbs sampling algorithm for a VAR (4)

- **Step 3 (continued)**: With the parameters of the inverse distribution to hand, namely $\overline{\Sigma} = \overline{S} + (Y_t - X_t B^1)'(Y_t - X_t B^1)$ and $T + \alpha$, we can use the above algorithm to draw Σ from the inverse Wishart distribution
- We repeat Steps 2 and 3 M times to obtain B^1 , B^2, ... , B^M and Σ^1, Σ^2, ... , Σ^M
- The last H values of B and Σ from these iterations is then used to form the empirical distribution of these parameters
- Note that the draws of the model parameters (after a burn-in period) are typically used to calculate forecasts or impulse response functions and build the distribution for the statistics of interest

Large BVARs

- Leeper *et al.* (1996) demonstrate that it is feasible to estimate VAR models with as many as 18 variables under the Sims and Zha prior
- With the advent of more powerful computers for Markov-chain Monte Carlo (MCMC) methodologies, including Gibbs sampling and the Metropolis-Hastings algorithm, much larger VARs can now be estimated by Bayesian methods
- The curse of dimensionality is handled using Bayesian shrinkage mechanisms, which have been shown to provide reliable estimates under general assumptions (De Mol *et al.* (2008), Bánbura *et al.* (2010))

Very large BVARs

- De Mol *et al.* (2008) assess the performance of BVARs for models with more than 100 variables
- They study forecasting accuracy and perform a structural exercise on the effect of a monetary policy shock on the macroeconomy
- The large BVAR outperforms smaller models in forecast accuracy and produces credible impulse responses, but this performance is already obtained with a 20-variable VAR
- Banbura *et al.* (2010) – as well as Koop (2013) – also show that medium-scale BVARs of about 20 variables deliver more accurate forecasts than large BVARs

Summary (1)

- VARs generally provide a credible approach to capturing the time-series properties of a vector time series as well as forecasting and structural inference
- But the number of coefficients in a VAR easily proliferates, putting limits on the number of variables that can be included in the VAR and on the precision of coefficient estimates
- One approach to addressing this curse of dimensionality is Bayesian VARs (BVARs)

Summary (2)

- Using Bayes' law, BVARs include prior information into the estimation process
- Early BVAR models used the Minnesota prior and the assumption that the error covariance matrix is fixed and diagonal
- Newer BVAR techniques using the Normal-inverse Wishart prior have removed this assumption

References and further reading (1)

Bańbura, M, Giannone, D and Reichlin, L (2010), 'Large Bayesian vector auto regressions', *Journal of Applied Econometrics*, Vol. 25, No. 1, pages 71-92.

Bauwens, L M, Lubrano, M and Richard, J F (1999), *Bayesian inference in dynamic econometric models*, Oxford, Oxford University Press.

Blake, A P and Mumtaz, H (2012), *Applied Bayesian econometrics for central bankers*, Centre for Central Banking Studies Technical Handbook - No. 4. http://www.bankofengland.co.uk/education/Documents/ccbs/technical_handbooks/pdf/techbook4.pdf.

Canova, F (2007), *Methods for applied macroeconomic research*, Princeton, Princeton University Press.

References and further reading (2)

Ciccarelli, M and Rebucci, A (2003), 'Bayesian VARs: a survey of the recent literature with an application to the European Monetary System', *IMF Working Paper WP/03/102*. http://www.imf.org/external/pubs/ft/wp/2003/wp03102.pdf.

De Mol, C, Giannone, D and Reichlin, L (2008), 'Forecasting using a large number of predictors: is Bayesian shrinkage a valid alternative to principal components?', *Journal of Econometrics*, Vol. 146, No. 2, pages 318-28.

Del Negro, M and Schorfheide, F (2004), 'Priors from general equilibrium models for VARs', *International Economic Review*, Vol. 45, No. 2, pages 643-73.

References and further reading (3)

Doan, T, Litterman, R and Sims, C A (1984), 'Forecasting and conditional projection using realistic prior distributions', *Econometric Reviews*, Vol. 3, No. 1, pages 1-100.

Kadiyala, K R and Karlsson, S (1997), 'Numerical methods for estimation and inference in Bayesian VAR-models', *Journal of Applied Econometrics*, Vol. 12, No. 2, pages 99-132.

Kim, C-J and Nelson, C R (1999), *State-space models with regime switching*, Cambridge, Mass, MIT Press.

Koop, G (2013), 'Forecasting with medium and large Bayesian VARs', *Journal of Applied Econometrics*, Vol. 28, No. 2, pages 177-203.

Koop, G, Poirier, D J and Tobias, J L (2007), *Bayesian econometric methods*, Cambridge, Cambridge University Press.

References and further reading (4)

Koop, G and Korobilis, D (2010), 'Bayesian multivariate time series methods for empirical macroeconomics', *Foundations and Trends in Econometrics*, Vol. 3, No. 4, pages 267-358. http://personal.strath.ac.uk/gary.koop/koop_korobilis_Foundations_and_Trends_2010.pdf.

Litterman, R (1986), 'Forecasting with Bayesian vector autoregressions – five years of experience', *Journal of Business & Economic Statistics*, Vol. 4, No. 1, pages 25-38.

Robertson, J C and Tallman, E W (1999), 'Vector autoregressions: forecasting and reality', *Federal Reserve Bank of Atlanta Economic Review*, Vol. 84, No. 1, pages 4-18. http://www.frbatlanta.org/filelegacydocs/robtallman.pdf.

References and further reading (5)

Sims, C S and Zha, T A (1998), 'Bayesian methods for dynamic multivariate models', *International Economic Review*, Vol. 39, No. 4, pages 949-68.

Sims, C S and Zha, T A (1999), 'Error bands for impulse responses', *Econometrica*, Vol. 67, No. 5, pages 1,113-56.

Waggoner, D F and Zha, T A (1999), 'Conditional forecasts in dynamic multivariate models', *Review of Economics and Statistics*, Vol. 80, No. 4, pages 639-51.

Zellner, A (1971), *An introduction to Bayesian inference in econometrics*, New York, Wiley.

Zha, T A (1998), 'A dynamic multivariate model for use in formulating policy', *Federal Reserve Bank of Atlanta Economic Review*, Vol. 83, No. 1, pages 16-29.