Bayesian Econometrics

Lecture 5: Bayesian VAR Models

Prof. Dr. Kai Carstensen

Kiel University

Winter Term 2024/25

Outline of this lecture

This lecture introduces to Bayesian VAR (BVAR) models.

1. Introduction
2. Some important multivariate distributions
3. Likelihood function
4. Priors: diffuse, Minnesota, natural conjugate, independent Normal-Wishart
5. Forecasting

Literature

Primary references:

- Koop and Korobilis (2010), Bayesian Multivariate Time Series Methods for Empirical Macroeconomics, Foundations and Trends in Econometrics, Vol. 3, No. 4, p. 267-358, downloadable here.
- Karlsson (2013), Forecasting with Bayesian Vector Autoregressions, in: Elliott and Timmermann (eds.), Handbook of Economic Forecasting, Vol. 2(B), p. 791-897, working paper version here.

Additional references:

- Kadiyala and Karlsson (1997), Numerical Methods for Estimation and Inference in Bayesian VAR-Models, Journal of Applied Econometrics 12(2), 99-132.
- Sims and Zha (1998), Bayesian Methods for Dynamic Multivariate Models, International Economic Review 39(4), 949-968.
- Banbura, Giannone and Reichlin (2010), Large Bayesian Vector Auto Regressions, Journal of Applied Econometrics 25(1), 71-92.
- Koop (2013), Forecasting with Medium and Large Bayesian VARs, Journal of Applied Econometrics 28, 177-203.

1. Introduction to VAR models

Model setup

VAR(p) model:

$$y_t = a_0 + A_1 y_{t-1} + \cdots + A_p y_{t-p} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma),$$

where

- y_t is a $M \times 1$ vector of time series variables,
- ε_t is a $M \times 1$ vector of reduced form errors,
- a_0 is a $M \times 1$ vector of intercepts,
- A_i is a $M \times M$ matrix of coefficients, and
- Σ is a $M \times M$ covariance matrix.

Defining the lag polynomial $A(L) = I - A_1 L - A_2 L^2 - \cdots - A_p L^p$, the VAR can be written as

$$A(L) y_t = a_0 + \varepsilon_t.$$

What do we need VAR models for?

Two major purposes:

- Structural economic analysis
- Forecasting

Structural economic analysis

Most prominent tool is impulse response analysis: how do the variables $y_t, y_{t+1}, y_{t+2}, \ldots$ react to a ceteris paribus shock to ε_t?

Since the ε_t's are correlated (with covariance matrix Σ), there is no ceteris paribus shock, say, ε_{jt}. This is why we call ε_t reduced form shocks.

Instead, we may assume that reduced form shocks are a linear combination of uncorrelated structural form shocks:

$$\varepsilon_t = De_t, \qquad e_t \sim \mathcal{N}(0, I_M).$$

Now let us suppose we can interpret e_{kt} as a structural monetary policy shock. A "typical" shock of size one standard deviation is a unit shock. It works as follows through the VAR model:

$$e_{kt} \to \varepsilon_t \to y_t \to y_{t+1} \to y_{t+2} \to \cdots$$

How can we find the parameters in D?

Note that Σ is a variance matrix of the VAR disturbances which can be estimated from the residuals.

Now, since

$$\Sigma = \mathrm{E}(\varepsilon_t \varepsilon_t') = \mathrm{E}(D e_t e_t' D') = D\,\mathrm{E}(e_t e_t') D' = DD',$$

we can derive D from Σ if we add $M(M-1)/2$ economically motivated *identifying restrictions*.

For details, see the course Macroeconometrics.

Vector moving average presentation

The impulse response analysis builds on the vector moving average presentation (VMA) of the VAR model:

$$y_t = A(L)^{-1}(a_0 + \varepsilon_t) = \mu + \sum_{i=0}^{\infty} C_i \varepsilon_{t-i} = \mu + C(L)\varepsilon_t = \mu + C(L)De_t,$$

where $\mu = A(L)^{-1}a_0 = A(1)^{-1}a_0$ and the matrices C_i are recursively calculated as

$$C_0 = I \qquad \text{for} \quad i = 0$$

$$C_i = \sum_{k=1}^{i} C_{i-k}A_k \qquad \text{for} \quad i = 1, 2, \ldots$$

To obtain these results, we require the VAR to be stable (=stationary). For details, see the course Multivariate Time Series Analysis and Forecasting.

The matrices $C_i D$ represent the partial effect of the structural shocks e_t on y_{t+i} since

$$y_{t+i} = C_0 D e_{t+i} + \ldots + C_i D e_t + \ldots = \tilde{C}_0 e_{t+i} + \ldots + \tilde{C}_i e_t + \ldots,$$

where $\tilde{C}_i = C_i D$.

Thus,

$$\frac{\partial y_{t+i}}{\partial \epsilon_t'} = \tilde{C}_i = \begin{pmatrix} \tilde{c}_{11}^i & \cdots & \tilde{c}_{1n}^i \\ \vdots & \ddots & \vdots \\ \tilde{c}_{n1}^i & \cdots & \tilde{c}_{nn}^i \end{pmatrix}$$

meaning that $\tilde{c}_{kl}^i$ is the partial effect of the l^{th} structural shock $e_{l,t}$ on the k^{th} variable i periods later, $y_{k,t+i}$.

The sequence $c_{kl}^0, c_{kl}^1, c_{kl}^2, \ldots$ is the impulse response function (IRF) of the k^{th} variable after structural shock l.

Forecasting

VAR models are a popular tool for forecasting.

Typical forecast quantities of interest: forecast mean and forecast intervals.

Using Bayesian techniques, we can of course find the whole forecast distribution.

More on this later.

Why is Bayesian estimation well-suited for VAR models?

VAR models by construction ("everything depends on everything") tend to be hugely parameterized.

For example, consider a typical monetary VAR model for the euro area estimated on quarterly data:

- Typically, $M = 5$ variables are included: output, prices, money, interest rates, exchange rates.
- To take reasonable adjustment lags into account, quarterly data require at least $p = 4$ lags (but better $p = 8$).
- $M = 5$, $p = 4$, 1 intercept: $Mp + 1 = 21$ parameters per VAR equation.
- $M = 5$, $p = 8$, 1 intercept: $Mp + 1 = 41$ parameters per VAR equation.
- Euro area data start in 1999Q1 which is why we currently have less than 80 observations.
- Huge parameter and forecast imprecision.

The Bayesian solution to overparameterization is effectively shrinkage:

- Typically, VAR priors are chosen which imply very simple models.
- Examples: y_t is white noise or y_t is a random walk.
- Structural perspective:
  - We shrink parameter estimates towards the prior means.
  - Sensible for structural analysis? Depends on whether the priors really reflect prior knowledge.
- Forecasting perspective:
  - We shrink forecasts towards forecasts of simple models.
  - Works often well because forecasting performance of simple models is good.
  - Hence, we may see the prior as an easy way to obtain shrinkage. We do not interpret it as carrying structural information.
  - In fact, forecasting with Bayesian VAR models has recently become increasingly popular because it allows to include many variables into a single model. For example, Banbura, Giannone and Reichlin (2010) include more than 130.
  - There are other shrinkage methods (ridge regression, LASSO, elastic net, boosting, ...).

2. Multivariate distributions

Wishart distribution

Let H be an $M \times M$ random matrix that follows a Wishart distribution with parameters Φ and ν,

$$H \sim W(\Phi, \nu).$$

Then it has pdf

$$f_W(H|\Phi, \nu) = c_W^{-1} |\Phi|^{-\frac{\nu}{2}} |H|^{\frac{\nu-M-1}{2}} \exp\left[-\frac{1}{2}\operatorname{tr}(\Phi^{-1}H)\right],$$

where c_W is an integration constant, $\nu > M-1$ is a scalar parameter, and Φ is an $M \times M$ symmetric and positive definite scale matrix.

The expectation is

$$\mathrm{E}(H) = \nu\Phi.$$

A good reference is Steven W. Nydick (2012), The Wishart and Inverse Wishart Distributions, downloadable here. Or look up the appendix in Karlsson (2013).

Inverse Wishart distribution

Let Σ be an $M \times M$ random matrix that follows an inverse Wishart distribution with parameters Ψ and δ,

$$\Sigma \sim iW(\Psi, \delta).$$

Then it has pdf

$$f_{iW}(\Sigma|\Psi, \delta) = c_{iW}^{-1}|\Psi|^{\frac{\delta}{2}}|\Sigma|^{-\frac{\delta+M+1}{2}} \exp\left[-\frac{1}{2}\operatorname{tr}(\Psi\Sigma^{-1})\right],$$

where c_{iW} is an integration constant, $\delta > M-1$ is a scalar parameter, and Ψ is an $M \times M$ symmetric and positive definite scale matrix.

Its expectation is

$$\mathrm{E}(\Sigma) = \Psi/(\delta - M - 1).$$

Relationship between Wishart and inverse Wishart

The Wishart distribution is a multivariate generalisation of the Gamma distribution. It is often used as a prior for the precision matrix (=inverse of the variance matrix).

The inverse Wishart distribution is a multivariate generalisation of the inverse Gamma distribution. It is often used as a prior for the variance matrix.

Relationship between Wishart and inverse Wishart distribution: Let

$$H \sim W(\Phi, \nu).$$

Then

$$\Sigma \equiv H^{-1} \sim iW(\Phi^{-1}, \nu).$$

Normal-Wishart and Normal-inverse Wishart distributions

The Normal-(inverse) Wishart distribution is a multivariate generalization of the Normal-(inverse) Gamma distribution.

If $\alpha|H \sim \mathcal{N}(\underline{\alpha}, H^{-1} \otimes \underline{V})$ and $H \sim W(\underline{S}^{-1}, \underline{\nu})$, then (α, H) is said to have Normal-Wishart distribution:

$$\alpha, H \sim NW(\underline{\alpha}, \underline{V}, \underline{S}^{-1}, \underline{\nu}).$$

If $\alpha|\Sigma \sim \mathcal{N}(\underline{\alpha}, \Sigma \otimes \underline{V})$ and $\Sigma \sim iW(\underline{S}, \underline{\nu})$, then (α, Σ) is said to have Normal-inverse Wishart distribution:

$$\alpha, \Sigma \sim NiW(\underline{\alpha}, \underline{V}, \underline{S}, \underline{\nu}).$$

Since for our purpose the two differ only by whether we parameterize our models with a precision matrix H or a variance matrix Σ, I will typically call it just Normal-Wishart even if I use the Normal-inverse Wishart. To my taste it is more straightforward to parameterize multivariate models in terms of the variance matrix.

Marginal distribution of Normal-inverse Wishart

Let $\alpha = \text{vec}(A)$, where A is a $(pM + 1) \times M$ dimensional random matrix. If $\alpha|\Sigma \sim \mathcal{N}(\underline{\alpha}, \Sigma \otimes \underline{V})$ and $\Sigma \sim iW(\underline{S}, \underline{\nu})$ such that they have joint Normal-inverse Wishart distribution

$$\alpha, \Sigma \sim NiW(\underline{\alpha}, \underline{V}, \underline{S}, \underline{\nu}),$$

then α has mean

$$\text{E}(\alpha) = \underline{\alpha}, \qquad \underline{\nu} > 1,$$

and variance

$$\text{Var}[\alpha] = \frac{\underline{S} \otimes \underline{V}}{\underline{\nu} - M - 1}, \qquad \underline{\nu} > M + 1.$$

The marginal distribution of A is analytical.

It can be shown that it is the matrix variate t distribution

$$A \sim MT(\underline{A}, \underline{V}, \underline{S}, \underline{\nu}),$$

see Dreze and Richard (1983), Bayesian analysis of simultaneous equation systems, in: Griliches and Intriligator (eds.), Handbook of Econometrics, Vol., Chapter 9, p. 517-598.

The matrix variate t distribution is also defined in Karlsson (2013) who also presents an algorithm to draw respective random numbers.

A nice feature of the matrix variate t distribution is that each element has a marginal univariate t distribution.

A single row i column j element A_{ij} has mean

$$\mathrm{E}(A_{ij}) = \underline{A}_{ij}, \qquad \underline{\nu} > 1,$$

and variance

$$\mathrm{Var}(A_{ij}) = \frac{\underline{V}_{ii}\underline{S}_{jj}}{\underline{\nu} - M - 1}, \qquad \underline{\nu} > M + 1.$$

Defining $\sigma_{ij}^2 = \underline{V}_{ii}\underline{S}_{jj}/(\underline{\nu} - M + 1)$, the standardized element

$$T_{ij} = \frac{A_{ij} - \underline{A}_{ij}}{\sigma_{ij}} = \frac{A_{ij} - \underline{A}_{ij}}{\sqrt{\underline{V}_{ii}\underline{S}_{jj}/(\underline{\nu} - M + 1)}}$$

has student t distribution with $\underline{\nu} - M + 1$ degrees of freedom,

$$T_{ij} \sim t(\underline{\nu} - M + 1)$$

which can be used to construct probability intervals. (If $\underline{\nu} - M + 1$ is large, then the student t distribution can of course be approximated well by the standard normal.)

3. The likelihood function

Bayesian estimation of the VAR

Suppose we want to estimate the parameters $a_0, A_1, \ldots, A_p, \Sigma$ with Bayesian methods.

Then we can proceed as always:

- Find the posterior which is proportional to the product of prior and likelihood.
- Compute interesting quantities (mean, variance, quantiles) of the posterior.

If we want to compute the IRF, just calculate (or simulate) the posterior distribution of $\tilde{C}_0, \tilde{C}_1, \tilde{C}_2, \ldots$. This is straightforward because the $\tilde{C}_i$'s are functions of $A_1, \ldots, A_p$ and Σ.

If we want to produce forecasts, we need to find the predictive density.

Let us start finding the likelihood function.

Representations of the VAR

For ease of notation, the transposed VAR(p) model

$$y_t' = a_0' + y_{t-1}' A_1' + \cdots + y_{t-p}' A_p' + \varepsilon_t'$$

can be written as

$$y_t' = [1, y_{t-1}', \ldots, y_{t-p}'] \begin{pmatrix} a_0' \\ A_1' \\ \vdots \\ A_p' \end{pmatrix} + \varepsilon_t' = x_t A + \varepsilon_t'$$

with the $(pM+1) \times M$ matrix A and the $1 \times (pM+1)$ vector x_t so defined. Stacking all observations $t = 1, \ldots, T$ yields the *matrix representation*

$$Y = XA + E,$$

where $Y = [y_1, \ldots, y_T]'$ is $T \times M$, $X = [x_1', \ldots, x_T']'$ is $T \times (pM+1)$, and $E = [\varepsilon_1, \ldots, \varepsilon_T]'$ is $T \times M$.

Note that E is not a vector which is why it has a matrix-variate normal distribution.

To obtain a vector-variate normal distribution (the one we know), we vectorize the system:

$$\text{vec}(Y) = \text{vec}(XA) + \text{vec}(E) = (I_M \otimes X)\,\text{vec}(A) + \text{vec}(E).$$

Defining $y = \text{vec}(Y)$, $\alpha = \text{vec}(A)$, $\mathbf{X} = (I_M \otimes X)$, and $\varepsilon = \text{vec}(E)$ yields the *vector representation*

$$y = \mathbf{X}\alpha + \varepsilon.$$

This looks like a regression equation. Thus, it is no surprise that the analysis of the VAR model is a kind of multivariate generalization of the regression analysis discussed earlier.

Properties of ε:

$$\varepsilon = [\varepsilon_{11}, \ldots, \varepsilon_{1T}, \varepsilon_{21}, \ldots, \varepsilon_{2T}, \ldots\ldots\ldots, \varepsilon_{M1}, \ldots, \varepsilon_{MT}]'$$

As shown in the tutorial, it has mean

$$\mathrm{E}(\varepsilon) = 0$$

and variance matrix

$$\mathrm{Var}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon') = \Sigma \otimes I_T.$$

Hence, due to the normality assumption,

$$\varepsilon \sim \mathcal{N}(0, \Sigma \otimes I_T).$$

The likelihood function

To obtain the likelihood function, we may use the conditional-marginal factorization of the joint pdf given past information $\mathcal{I}_{t-1}$:

$$f(y|y_{1-p},\ldots,y_0) = f(y_T|\mathcal{I}_{T-1}) \cdot f(y_{T-1}|\mathcal{I}_{T-2}) \cdot \ldots \cdot f(y_1|\mathcal{I}_0),$$

where we condition on p pre-sample values.

Since the conditional distribution of y_t is normal with

$$\mathsf{E}(y_t|\mathcal{I}_{t-1}) = a_0 + A_1 y_{t-1} + \cdots + A_p y_{t-p} = A' x_t'$$

and

$$\mathsf{Var}(y_t|\mathcal{I}_{t-1}) = \mathsf{Var}(\varepsilon_t) = \Sigma$$

we obtain a normal pdf.

The conditional period-t pdf is (leaving out constants)

$$f(y_t|\mathcal{I}_{t-1}) = |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y_t - A'x_t')'\Sigma^{-1}(y_t - A'x_t')\right]$$

This gives rise to the joint density

$$f(y|A,\Sigma) = |\Sigma|^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - A'x_t')'\Sigma^{-1}(y_t - A'x_t')\right],$$

where we leave out the pre-sample values $y_{1-p}, \ldots, y_0$ from the conditioning set and add instead the parameters A and Σ to stress that the pdf will be interpreted as likelihood function.

As shown in the tutorial, the pdf can be re-written using either the matrix representation of the VAR:

$$f(y|\alpha,\Sigma) \propto |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\,\mathrm{tr}\left[(Y-XA)'(Y-XA)\Sigma^{-1}\right]\right\}$$

or its vector representation:

$$f(y|\alpha,\Sigma) \propto |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}(y-\mathbf{X}\alpha)'(\Sigma^{-1}\otimes I_T)(y-\mathbf{X}\alpha)\right\}.$$

How can we easily transform the the two representations of the likelihood function into each other?

We need the rule: $\text{tr}(PQRS) = \text{vec}(P')'(S' \otimes Q)\,\text{vec}(R)$ for matrices P, Q, R, S of appropriate dimensions.

Now first recall that $y - \mathbf{X}\alpha = \text{vec}(Y - XA)$. Hence,

$$(y - \mathbf{X}\alpha)'(\Sigma^{-1} \otimes I_T)(y - \mathbf{X}\alpha) = \text{vec}(Y - XA)'(\Sigma^{-1} \otimes I_T)\,\text{vec}(Y - XA).$$

Apply the rule:

$$\text{vec}(\underbrace{Y - XA}_{P'})'(\underbrace{\Sigma^{-1}}_{S'} \otimes \underbrace{I_T}_{Q})\,\text{vec}(\underbrace{Y - XA}_{R}) = \text{tr}\left[\underbrace{(Y - XA)'}_{P}\underbrace{(Y - XA)}_{R}\underbrace{\Sigma^{-1}}_{S}\right]$$

Normal-Wishart structure of the likelihood

As shown in the tutorial, the likelihood function of the VAR(p) model with M variables can be rearranged to

$$f(y|\alpha,\Sigma) \propto |\Sigma|^{-\frac{pM+1}{2}} \exp\left[-\frac{1}{2}(\alpha-\hat{\alpha})'(\Sigma^{-1}\otimes X'X)(\alpha-\hat{\alpha})\right]$$

$$|\Sigma|^{-\frac{\nu+M+1}{2}} \exp\left[-\frac{1}{2}\operatorname{tr}(S\Sigma^{-1})\right]$$

which is the kernel of a Normal-Wishart distribution

$$\alpha|\Sigma \sim \mathcal{N}(\hat{\alpha}, \Sigma\otimes(X'X)^{-1})$$

and

$$\Sigma \sim iW(S,\nu)$$

with $S=(Y-X\hat{A})'(Y-X\hat{A})$, $\nu = T-pM-M-2$, $\hat{A}=(X'X)^{-1}X'Y$ and $\hat{\alpha}=\text{vec}(\hat{A})$.

The Normal-Wishart structure of the likelihood implies that a natural conjugate prior for α and Σ also needs to have this structure. In fact, the natural conjugate prior is

$$\alpha|\Sigma \sim \mathcal{N}(\underline{\alpha}, \Sigma \otimes \underline{V})$$

and

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

where $\underline{\alpha}$, $\underline{V}$, $\underline{S}$, $\underline{\nu}$ are prior hyperparameters.

We come back to this point later.

4. Diffuse prior

Prior

$$f(\alpha, \Sigma) \propto |\Sigma|^{-\frac{M+1}{2}}$$

Discussion:

- This is an improper prior paralleling the one we used in the Normal-Gamma regression model.
- Typically, it is not advisable because it does not allow for shrinkage.

Posterior

As shown in the tutorial, the posterior is Normal-Wishart:

$$\alpha|\Sigma, y \sim \mathcal{N}(\hat{\alpha}, \Sigma \otimes (X'X)^{-1})$$

and

$$\Sigma|y \sim iW([Y - X\hat{A}]'[Y - X\hat{A}], T - pM - 1)$$

where $\hat{A} = (X'X)^{-1}X'Y$ and $\hat{\alpha} = \text{vec}(\hat{A})$.

Hence, the posterior mean

$$\text{E}(A|y) = \hat{A}$$

is the OLS estimator.

5. Minnesota prior

Prior

Take $\Sigma = \hat{\Sigma}$ as given. This is not very Bayesian because we cannot integrate it out.

For α use a normal prior: $\alpha \sim \mathcal{N}(\underline{\alpha}, \underline{V}_M)$.

Advantage:

- Likelihood has, conditional on $\Sigma = \hat{\Sigma}$, a normal structure (with mean $\hat{\alpha}$ and variance $\hat{\Sigma} \otimes (X'X)^{-1}$).
- Prior for α is normal.
- Thus, posterior for α is normal. Nice and simple!

Posterior

As shown in the tutorial, the posterior is

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

with

$$\bar{V}_M = (\hat{\Sigma}^{-1} \otimes X'X + \underline{V}_M^{-1})^{-1}$$

and

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

Hence, posterior inference is easy.

Specifying the prior

While there are, in principle, many ways to implement the prior assumptions, the literature often uses a specific "Minnesota" specification.

Estimation of Σ:

- Σ is assumed to be diagonal.
- Its diagonal elements σ_{ii} are estimated as residual variances of pth order autoregressions for variable i.

Idea for prior specification:

- Persistent variables: prior is pure random walk.
- Non-persistent variables: prior is white noise.

Prior mean for $A_1, \ldots, A_p$ of persistent variables (let a_{ij}^k be the row i column j element of A_k):

$$\mathrm{E}(a_{ij}^k) = \begin{cases} 1 & \text{if } i = j \text{ and } k = 1 \text{ (diagonal elements of first lag)} \\ 0 & \text{otherwise} \end{cases}$$

Prior mean for $A_1, \ldots, A_p$ of non-persistent variables:

$$\mathrm{E}(a_{ij}^k) = \begin{cases} 0 & \text{if } i = j \text{ and } k = 1 \\ 0 & \text{otherwise} \end{cases}$$

Prior mean for intercept (let a_i^0 be the row i element of a_0):

$$\mathrm{E}(a_i^0) = 0$$

$\rightarrow$ Prior variance for $A_1, \ldots, A_p$ (let σ_{ij} be the row i column j element of Σ):

$$\text{Var}(a_{ij}^k) = \begin{cases} \frac{\lambda^2}{k^2} & \text{if } i = j \\ \theta_1^2 \frac{\lambda^2}{k^2} \frac{\sigma_{ii}}{\sigma_{jj}} & \text{otherwise} \end{cases}$$

$\rightarrow$ Prior variance for intercept:

$$\text{Var}(a_i^0) = \theta_2^2 \lambda^2 \sigma_{ii}$$

$\rightarrow$ All prior covariances are set to zero.

Interpretation (note a slightly different parameterization in Koop and Korobilis, 2010):

- $\lambda > 0$ controls overall tightness of the prior
- $\theta_1 > 0$ controls deviating tightness of prior for nondiagonal elements
- $\theta_2 > 0$ controls deviating tightness of prior for intercept
- factor $1/k^2$ increases tightness (around mean zero) for higher lags
- factor σ_{ii}/σ_{jj} accounts for differences in scale between variables i and j
- factor σ_{ii} accounts for scale in intercept tightness

Numerical interpretation of λ:

- λ affects the variance quadratically and thus the s.d. linearly.
- Hence, doubling λ increases the prior 95% interval for any parameter by factor 2.

Numerical interpretation of θ_1:

- θ_1 affects the variance of the off-diagonal elements in A_k quadratically and thus their s.d. linearly.
- Hence, $\theta_1 = 0.5$ ceteris paribus renders the prior length of an 95% interval for off-diagonal parameters half as long as for diagonal parameters.
- θ_1 is often chosen to be smaller than one to reflect our prior knowledge, that own lags are more important which is why we put a less strict prior on own lags an thus allow the data to affect the posterior more.

Numerical interpretation of θ_2:

- θ_2 affects the variance of the elements in a_0 quadratically and thus their s.d. linearly.
- Hence, $\theta_2 = 2$ ceteris paribus renders the prior length of an 95% interval for intercept elements twice as long as for diagonal parameters.
- θ_2 is often chosen to be very large to effectively impose a diffuse prior.

Numerical interpretation of factor $1/k^2$:

- $1/k$ affects the variance of the elements in A_k, $k = 1, \ldots, p$, quadratically and thus their s.d. linearly.
- Denote the prior length of an 95% interval for elements in A_1 by Δ.
- The prior length of an 95% interval for elements in $A_2, \ldots, A_p$ are thus $\frac{1}{2}\Delta, \ldots, \frac{1}{p}\Delta$.
- This reflects our prior knowledge that higher lags are less important and should thus be shrunk more towards zero.
- Note that some authors let the variance shrink with $1/k$ and thus the s.d. and interval length with $1/\sqrt{k}$.

Numerical interpretation of factor σ_{ii}/σ_{jj}:

- Note the interpretation $a_{ij}^k = \partial y_{it}/\partial y_{jt-k}$, keeping everything else fixed.
- Hence, changing the scale of, say y_{it}, from Billions to Millions increases a_{ij}^k by a factor of 1000.
- To ensure that the scale of a prior 95% interval changes accordingly, we require a quadratic correction factor such as the variance of y_{it} or, simpler, the variance of the shock ε_{it} which is what the literature does.
- For the same reasoning, the variance of the shock ε_{jt} is included in the denominator.

Numerical interpretation of factor σ_{ii} for intercept tightness:

- Same reasoning as above.

Implementation:

- Since σ_{ij} are unknown, use estimates $\hat{\sigma}_{ij}$ (see above).

6. Natural conjugate prior

Prior

Given the structure of the likelihood, the natural conjugate prior is

$$\alpha|\Sigma \sim \mathcal{N}(\underline{\alpha}, \Sigma \otimes \underline{V})$$

and

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

where $\underline{\alpha}$, $\underline{V}$, $\underline{S}$, $\underline{\nu}$ are prior hyperparameters.

Comparison with the Minnesota prior:

- Σ is treated as random variable instead of taking some (crude) estimator.
- The prior for Σ is a straightforward generalization of the gamma.
- The prior for α differs from the Minnesota prior in an important respect: $\text{Var}(\alpha|\Sigma)$ has the structure $\Sigma \otimes \underline{V}$. This has some consequences for the specification (see below).

Posterior

The posterior is Normal-Wishart:

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

and

$$\Sigma|y \sim iW(\bar{S}, \bar{\nu}),$$

where

$$\bar{V} = (\underline{V}^{-1} + X'X)^{-1}$$

$$\bar{\alpha} = \text{vec}(\bar{A}), \qquad \bar{A} = \bar{V}(\underline{V}^{-1}\underline{A} + X'X\hat{A})$$

$$\bar{S} = \underline{S} + \underline{A}'\underline{V}^{-1}\underline{A} + Y'Y - \bar{A}'\bar{V}^{-1}\bar{A}$$

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

Marginal posterior of α

As $\alpha = \text{vec}(A)$ is from a Normal-Wishart distribution, recall that the marginal distribution of A is matrix variate t:

$$A \sim MT(\bar{A}, \bar{V}, \bar{S}, \bar{\nu}).$$

A single row i column j element A_{ij} has a (nonstandardized) t distribution which is why the standardized element

$$T_{ij} = \frac{A_{ij} - \bar{A}_{ij}}{\sigma_{ij}} = \frac{A_{ij} - \bar{A}_{ij}}{\sqrt{\bar{V}_{ii}\bar{S}_{jj}/(\bar{\nu} - M + 1)}}$$

has student t distribution with $\bar{\nu} - M + 1$ degrees of freedom,

$$T_{ij} \sim t(\bar{\nu} - M + 1)$$

which can be used to construct probability intervals. (If $\bar{\nu} - M + 1$ is large, then the student t distribution can of course be approximated well by the standard normal.)

Fictitious sample interpretation

Comparison of the likelihood

$$\alpha|\Sigma \sim \mathcal{N}(\hat{\alpha}, \Sigma \otimes (X'X)^{-1}), \qquad \Sigma \sim iW((Y - X\hat{A})'(Y - X\hat{A}), T - pM - M - 2)$$

and the prior

$$\alpha|\Sigma \sim \mathcal{N}(\underline{\alpha}, \Sigma \otimes \underline{V}), \qquad \Sigma \sim iW(\underline{S}, \underline{\nu}),$$

shows that the prior can be given a fictitious prior sample interpretation.

Suppose there is a fictitious prior sample Y_0 and X_0 of size T_0 with $\hat{A}_0 = (X_0'X_0)^{-1}X_0'Y_0$. Then one may set

$$\underline{\alpha} = \text{vec}(\hat{A}_0)$$
$$\underline{V} = (X_0'X_0)^{-1}$$
$$\underline{S} = (Y_0 - X_0\hat{A}_0)'(Y_0 - X_0\hat{A}_0)$$
$$\underline{\nu} = T_0 - pM - M - 2$$

Specifying the prior

We need to specify priors for both α and Σ. Here is a possible way to proceed:

- $\mathrm{E}(\alpha) = \underline{\alpha}$ can be specified as for the Minnesota prior (or use prior economic knowledge)
- $\mathrm{Var}(\alpha) = (\underline{S} \otimes \underline{V})/(\underline{\nu} - M - 1)$ is more restrictive than the Minnesota prior (see next page). Often, we would set only the diagonal elements to positive values that reflect how certain we are regarding $\underline{\alpha}$, and set the covariances to zero.
- $\underline{\nu}$ may be thought of a parameter that reflects the overall information content of our prior. Recall that in a fictitious prior sample interpretation, we would set it as $\underline{\nu} = T_0 - pM - M - 2$. Hence, a larger value of $\underline{\nu}$ corresponds to a larger fictitious prior sample.

Variance structure of the prior for α

The natural conjugate prior has a more restrictive prior variance structure,

$$\text{Var}(\alpha) = (\underline{S} \otimes \underline{V})/(\underline{\nu} - M - 1),$$

than, e.g., the Minnesota prior (or the independent Normal-Wishart prior),

$$\text{Var}(\alpha) = \underline{V}_M.$$

Count the elements (taking the symmetric structure into account):

- $\underline{S} \otimes \underline{V}$ is a $(pM^2 + M) \times (pM^2 + M)$ matrix that consists of the $M(M+1)/2$ distinct elements of $\underline{S}$ plus the $(pM+1)(pM+2)/2$ elements of $\underline{V}$.
- $\underline{V}_M$ is a $(pM^2 + M) \times (pM^2 + M)$ matrix that consists of $(pM^2 + M)(pM^2 + M + 1)/2$ distinct elements. This is much more.

Example: VAR(1) for $y_t = (gdp_t, m_t)'$. For simplicity no intercept.

$$\begin{pmatrix} gdp_t \\ m_t \end{pmatrix} = \begin{bmatrix} a_{11}^1 & a_{12}^1 \\ a_{21}^1 & a_{22}^1 \end{bmatrix} \begin{pmatrix} gdp_{t-1} \\ m_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}$$

Prior knowledge $\mathrm{E}(\alpha) = \underline{\alpha} = (1, 0, 0, 1)'$:

(1) gdp_t is persistent, hence we set $\underline{a}_{11}^1 = 1$.

(2) m_t is persistent, hence we set $\underline{a}_{22}^1 = 1$.

(3) gdp_t is not affected by m_t (perfect monetary neutrality), hence we set $\underline{a}_{12}^1 = 0$.

(4) m_t might be affected by gdp_t but sign unknown, hence we set $\underline{a}_{21}^1 = 0$.

How confident are we of our prior knowledge, i.e, how should we choose $\mathrm{Var}(\alpha)$?

(a) We are pretty sure of (1) to (3), so we want to set a tight prior (small prior variance).

(b) We are unsure concerning (4), so we want to set a loose prior (high prior variance).

$$\text{Var}(\alpha) = \frac{\underline{S} \otimes \underline{V}}{\underline{\nu} - 3} = \frac{\begin{bmatrix} \underline{S}_{11}\underline{V} & \underline{S}_{12}\underline{V} \\ \underline{S}_{21}\underline{V} & \underline{S}_{22}\underline{V} \end{bmatrix}}{\underline{\nu} - 3} = \frac{\begin{bmatrix} \underline{S}_{11}\underline{V}_{11} & \underline{S}_{11}\underline{V}_{12} & \underline{S}_{12}\underline{V}_{11} & \underline{S}_{12}\underline{V}_{12} \\ \underline{S}_{11}\underline{V}_{21} & \underline{S}_{11}\underline{V}_{22} & \underline{S}_{12}\underline{V}_{21} & \underline{S}_{12}\underline{V}_{22} \\ \underline{S}_{21}\underline{V}_{11} & \underline{S}_{21}\underline{V}_{12} & \underline{S}_{22}\underline{V}_{11} & \underline{S}_{22}\underline{V}_{12} \\ \underline{S}_{21}\underline{V}_{21} & \underline{S}_{21}\underline{V}_{22} & \underline{S}_{22}\underline{V}_{21} & \underline{S}_{22}\underline{V}_{22} \end{bmatrix}}{\underline{\nu} - 3}$$

Suppose a prior variance of 0.1 is tight. Then we set

(1) $\text{Var}(\underline{a}_{11}^1) = \text{Var}(\alpha_1) = \underline{S}_{11}\underline{V}_{11}/(\underline{\nu} - 3) = 0.1$

(2) $\text{Var}(\underline{a}_{12}^1) = \text{Var}(\alpha_2) = \underline{S}_{11}\underline{V}_{22}/(\underline{\nu} - 3) = 0.1$

(3) $\text{Var}(\underline{a}_{22}^1) = \text{Var}(\alpha_4) = \underline{S}_{22}\underline{V}_{22}/(\underline{\nu} - 3) = 0.1$

Note that (1) and (2) implies $\underline{V}_{11}/\underline{V}_{22} = 1 \quad \Rightarrow \quad \underline{V}_{11} = \underline{V}_{22}$.

This implies that the prior variance for the effect of gdp_t on m_t is also tight (hence we cannot set it to be loose):

(4) $\text{Var}(\underline{a}_{21}^1) = \text{Var}(\alpha_3) = \underline{S}_{22}\underline{V}_{11}/(\underline{\nu} - 3) = \underline{S}_{22}\underline{V}_{22}/(\underline{\nu} - 3) = \text{Var}(\alpha_4) = 0.1$

Using DSGE models to specify the priors

Given the fictitious prior sample interpretation of the natural conjugate prior, we may use economic models such as DSGE macromodels to find reasonable priors.

Simplest approach: Calibrate a DSGE model with fixed deep parameter values, simulate the model to obtain a "prior sample", use this to set the prior parameters. Choose the size T_0 of the prior sample according to how much weight you want to place on the prior.

Better: Choose prior distributions for the DSGE model's deep parameters, derive or simulate the implied prior distribution for the VAR parameters, proceed as above.

Reference: DelNegro and Schorfheide (2004) Priors from general equilibrium models for VARs, International Economic Review, 45(2), 643-673.

7. Independent Normal-Wishart prior

Prior

Priors for α and Σ are independent of each other.

Prior for α:

$$\alpha \sim \mathcal{N}(\underline{\alpha}, \underline{W})$$

Prior for Σ:

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

Discussion:

- Prior for α like the Minnesota one.
- Prior specification $\underline{\alpha}$ and $\underline{W}$ thus follows the same ideas.
- Prior for Σ flexible, unlike Minnesota one.

Posterior

Posterior pdf:

$$f(\alpha, \Sigma|y) \propto |\Sigma|^{-\frac{T+\underline{\nu}+M+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[(\underline{S} + Y'Y)\Sigma^{-1}\right]\right\}$$
$$\times \exp\left\{-\frac{1}{2}\left[\alpha'(\underline{W}^{-1} + \Sigma^{-1} \otimes X'X)\alpha - 2[\underline{\alpha}'\underline{W}^{-1} + y'(\Sigma^{-1} \otimes X)]\alpha\right]\right\}$$

This not a known pdf kernel.

Fortunately, it is possible to find known conditional distributions $f(\alpha|\Sigma, y)$ and $f(\Sigma|\alpha, y)$

The posterior pdf can be factorized into

$$f(\alpha, \Sigma|y) = f(\alpha|\Sigma, y)f(\Sigma|y),$$

where $f(\alpha|\Sigma, y)$ is a normal distribution with parameters

$$\bar{\alpha} = \bar{W}\left[\underline{W}^{-1}\underline{\alpha} + (\Sigma^{-1} \otimes X')y\right] \quad \text{and} \quad \bar{W} = (\underline{W}^{-1} + \Sigma^{-1} \otimes X'X)^{-1}.$$

Alternatively, it can be factorized into

$$f(\alpha, \Sigma|y) = f(\Sigma|\alpha, y)f(\alpha|y),$$

where $f(\Sigma|\alpha, y)$ is an inverse Wishart distribution with parameters

$$\bar{S} = \underline{S} + (Y - XA)'(Y - XA) \quad \text{and} \quad \bar{\nu} = T + \underline{\nu}.$$

Based on the conditionals, draws from the posterior can be obtained by a Gibbs sampler.

8. Forecasting

Setup

Aim:

- Forecast $y_{T+1:T+H} = (y'_{T+1}, \ldots, y'_{T+H})'$

Sample:

- Observations y_1 to y_T
- To be precise regarding our conditioning set, let us define $\mathbf{Y}_T = \{y_t\}_{t=1}^T$.
- Like before, for notational simplicity we neglect our conditioning on p pre-sample observations.

Parameters:

- Denote the vector of all parameters by θ.
- Minnesota prior: $\theta = \alpha$
- Natural conjugate prior: $\theta = (\alpha', \text{vec}(\Sigma)')'$

Predictive density

Typical quantities of interest:

- Average (optimal predictor under mean squared error loss)
- Median (optimal predictor under mean absolute error loss)
- Quantiles (often 5% and 95%)

Write them as

$$\mathrm{E}[g(y_{T+1:T+H})|\mathbf{Y}_T] = \int g(y_{T+1:T+H})p(y_{T+1:T+H}|\mathbf{Y}_T)dy_{T+1:T+H}$$

Predictive density:

$$p(y_{T+1:T+H}|\mathbf{Y}_T) = \int \underbrace{p(y_{T+1:T+H}|\mathbf{Y}_T,\theta)}_{\text{out-of-sample likelihood}} \underbrace{p(\theta|\mathbf{Y}_T)}_{\text{posterior}} d\theta$$

Sampling from the predictive density

Generally, the predictive density does not have an analytic form. (Exception: 1-step prediction under Minnesota or natural conjugate prior leads to predictive multivariate t distribution.)

Thus, we need to use Monte Carlo techniques.

Step 1:

- Sample θ from the posterior.
- Depending on the posterior, this can be done using the normal distribution, Gibbs sampling etc.

Step 2:

- Given θ, sample $y_{T+1:T+H}$.
- To this end, sample $\varepsilon_{T+i} \sim \mathcal{N}(0, \Sigma)$ for $i = 1, \ldots, H$.
- Use the VAR to recursively compute $y_{T+1}, \ldots, y_{T+H}$.

Sampling $y_{T+1}, \ldots, y_{T+H}$

(1) Take a draw $\theta^{(s)}$, and thus $\alpha^{(s)}$ and $\Sigma^{(s)}$, from the posterior.

(2) Draw $\varepsilon_{T+1}^{(s)}, \ldots, \varepsilon_{T+H}^{(s)}$ independently from the $\mathcal{N}(0, \Sigma^{(s)})$ distribution.

(3) Recursively compute $y_{T+1}^{(s)}, \ldots, y_{T+H}^{(s)}$:

$$
\begin{aligned}
y_{T+1}^{(s)} &= a_0^{(s)} + A_1^{(s)} y_T + A_2^{(s)} y_{T-1} + A_3^{(s)} y_{T-2} + \cdots + A_p^{(s)} y_{T-p+1} + \varepsilon_{T+1}^{(s)} \\
y_{T+2}^{(s)} &= a_0^{(s)} + A_1^{(s)} y_{T+1}^{(s)} + A_2^{(s)} y_T + A_3^{(s)} y_{T-1} + \cdots + A_p^{(s)} y_{T-p+2} + \varepsilon_{T+2}^{(s)} \\
y_{T+3}^{(s)} &= a_0^{(s)} + A_1^{(s)} y_{T+2}^{(s)} + A_2^{(s)} y_{T+1}^{(s)} + A_3^{(s)} y_T + \cdots + A_p^{(s)} y_{T-p+3} + \varepsilon_{T+3}^{(s)} \\
&\vdots
\end{aligned}
$$

(4) Repeat (1) to (3) $S = S_0 + S_1$ times. Discard the first S_0 burn-in replications.

Computing quantities of interest

For general function $g(y_{T+1:T+H})$:

$$\hat{g}(y_{T+1:T+H}) \equiv \frac{1}{S_1} \sum_{s=S_0+1}^{S} g(y_{T+1:T+H}^{(s)}) \xrightarrow{\text{p}} \mathsf{E}[g(y_{T+1:T+H})|\mathbf{Y}_T]$$

Example: predictive mean

$$\hat{y}_{T+i} \equiv \frac{1}{S_1} \sum_{s=S_0+1}^{S} y_{T+i}^{(s)}, \quad i = 1, \ldots, H$$

9. Forecasting with large BVARs

Banbura, Giannone and Reichlin (2010), Large Bayesian Vector Auto Regressions, Journal of Applied Econometrics 25(1), 71-92.

Introduction

Big questions:

- Are large BVARs useful for forecasting? (We will concentrate on this part.)
- Do they produce sensible results in structural analyses?

What the paper does (in terms of forecasting):

- Estimate BVARs with 3, 7, 20, and 131 variables.
- Make the amount of shrinkage data dependent.
- Evaluate the MSFEs of the BVARs relative to a benchmark.

VAR models

BGR consider VAR(p) models with the following variables:

- SMALL: employment (EMPL), CPI, Federal Funds rate (FFR)
- CEE: additionally commodity prices, non-borrowed reserves, total reserves, M2 money stock (this is the set of variables used in the monetary VAR of Christiano, Eichenbaum, Evans, 1999)
- MEDIUM: additionally personal income, real consumption, industrial production, capacity utilization, unemployment rate, housing starts, producer price index, personal consumption expenditures price deflator, average hourly earnings, M1 money stock, S&P stock price index, 10 year US Treasury Bond yields, effective exchange rate.
- LARGE: all 131 variables used in Stock and Watson (2005)

Prior

BGR start with a natural conjugate Normal-Wishart prior

$$\alpha|\Sigma \sim \mathcal{N}(\underline{\alpha}, \Sigma \otimes \underline{V})$$

and

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

Fictitious sample: dummy observations

BGR use the fictitious sample interpretation of the prior and define $T_0 = pM + M + 1$ dummy observations Y_0 and X_0 with the properties

$$\underline{\alpha} = \text{vec}(\hat{A}_0) = \text{vec}[(X_0'X_0)^{-1}X_0'Y_0]$$
$$\underline{V} = (X_0'X_0)^{-1}$$
$$\underline{S} = (Y_0 - X_0\hat{A}_0)'(Y_0 - X_0\hat{A}_0)$$

They also set

$$\underline{\nu} = T_0 - pM - 1 = M.$$

See below for the exact definition of Y_0 and X_0.

Posterior

They obtain the posterior

$$\alpha|y, \Sigma \sim \mathcal{N}(\bar{\alpha}, \Sigma \otimes \bar{V}), \quad \Sigma|y \sim iW(\bar{S}, \bar{\nu}),$$

where

$$\bar{V} = (\underline{V}^{-1} + X'X)^{-1}$$

$$\bar{\alpha} = \text{vec}(\bar{A}), \qquad \bar{A} = \bar{V}(\underline{V}^{-1}\underline{A} + X'X\hat{A})$$

$$\bar{S} = \underline{S} + \underline{A}'\underline{V}^{-1}\underline{A} + Y'Y - \bar{A}'\bar{V}^{-1}\bar{A}$$

$$\bar{\nu} = T + \underline{\nu} = T + M$$

Augmented VAR model

Augmenting the variables as

$$Y_* = \begin{pmatrix} Y \\ Y_0 \end{pmatrix}, \quad X_* = \begin{pmatrix} X \\ X_0 \end{pmatrix}, \quad E_* = \begin{pmatrix} E \\ E_0 \end{pmatrix}$$

BGR consider the augmented VAR model with $T_* = T + T_0$ observations:

$$Y_* = X_* A + E_*.$$

It is straightforward to show that the posterior quantities can be expressed as functions of the augmented model:

$$\bar{V} = (\underline{V}^{-1} + X'X)^{-1} = (X_0'X_0 + X'X)^{-1} = (X_*'X_*)^{-1}$$

$$\bar{A} = \bar{V}(\underline{V}^{-1}\underline{A} + X'X\hat{A}) = (X_*'X_*)^{-1}(X_0'X_0\hat{A}_0 + X'X\hat{A}) = (X_*'X_*)^{-1}X_*'Y_*$$

$$\bar{S} = \underline{S} + \underline{A}'\underline{V}^{-1}\underline{A} + Y'Y - \bar{A}'\bar{V}^{-1}\bar{A} = (Y_* - X_*\bar{A})'(Y_* - X_*\bar{A})$$

Estimation of the augmented VAR with diffuse prior

Estimating the augmented VAR with a diffuse prior,

$$f(\alpha, \Sigma) \propto |\Sigma|^{-\frac{M+1}{2}}$$

leads to the Normal-Wishart posterior (see above)

$$\alpha|\Sigma, y \sim \mathcal{N}(\bar{\alpha}_*, \Sigma \otimes \bar{V}_*), \qquad \Sigma|y \sim iW(\bar{S}_*, \bar{\nu}_*)$$

where

$$\bar{V}_* = (X_*'X_*)^{-1}$$

$$\bar{A}_* = (X_*'X_*)^{-1}X_*'Y_*$$

$$\bar{S}_* = (Y_* - X_*\bar{A}_*)'(Y_* - X_*\bar{A}_*)$$

$$\bar{\nu}_* = T_* - pM - 1 = T + T_0 - pM - 1 = T + M$$

Now compare the posterior quantities of the augmented VAR model estimated with diffuse prior (see p. 72) and the posterior quantities of the baseline VAR model estimated with Normal-Wishart prior (see p. 70).

It turns out they are the same:

$$\bar{V}_* = \bar{V}$$

$$\bar{A}_* = \bar{A}$$

$$\bar{S}_* = \bar{S}$$

$$\bar{\nu}_* = \bar{\nu}$$

A slightly different prior

BGR actually use a slightly different prior.

Reason: There is a problem with the Wishart prior $\Sigma \sim iW(\underline{S}, \underline{\nu})$ if $\underline{\nu} = M$: Its mean

$$\mathrm{E}(\Sigma) = \underline{S}/(\underline{\nu} - M - 1)$$

does not exist. It only exists if $\underline{\nu} > M + 1$.

To ensure existence, BGR set $\underline{\nu} = M + 2$ in the normal Wishart prior of the original VAR. To obtain the same results with the augmented VAR, they use the improper prior

$$f(\alpha, \Sigma) \propto |\Sigma|^{-\frac{M+3}{2}},$$

which slightly differs from the diffuse prior. This choice lets $\bar{V} = \bar{V}_*$, $\bar{A} = \bar{A}_*$, and $\bar{S} = \bar{S}_*$ unchanged but leads to $\bar{\nu} = \bar{\nu}_* = T + M + 2$.

Specification of the prior

The prior is very similar to the Minnesota prior discussed above.

Prior mean for $A_1, \ldots, A_p$ ($\delta_i = 1$ for persistent variables and zero otherwise):

$$\mathrm{E}(a_{ij}^k) = \begin{cases} \delta_i & \text{if } i = j \text{ and } k = 1 \text{ (diagonal elements of first lag)} \\ 0 & \text{otherwise} \end{cases}$$

Prior mean for intercept:

$$\mathrm{E}(a_i^0) = 0$$

Prior mean for variance:

$$\mathrm{E}(\Sigma) = \hat{\Sigma}_M = \mathrm{diag}([\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{MM}]),$$

where $\hat{\sigma}_{ii}$ is the residual variance of a pth order autoregression for variable i.

Prior variance for $A_1, \ldots, A_p$ (let σ_{ij} be the row i column j element of Σ):

$$\text{Var}(a_{ij}^k) = \left\{ \begin{array}{ll} \frac{\lambda^2}{k^2} & \text{if } i = j \\ \frac{\lambda^2}{k^2} \frac{\hat{\sigma}_{ii}}{\hat{\sigma}_{jj}} & \text{otherwise} \end{array} \right\} = \frac{\lambda^2}{k^2} \frac{\hat{\sigma}_{ii}}{\hat{\sigma}_{jj}}$$

Prior variance for intercept (BGR choose a very large θ_2 to obtain an effectively diffuse prior):

$$\text{Var}(a_i^0) = \theta_2^2 \lambda^2 \hat{\sigma}_{ii}$$

All prior covariances are set to zero.

Implementation of the priors

How to determine $\underline{\alpha}$

Question to be answered: how are the elements $\underline{a}_i^0$ of the intercept $\underline{a}_0$ and $\underline{a}_{ij}^k$ of the matrices $\underline{A}_k$ ordered in $\underline{\alpha}$?

Start with $\underline{A}$:

$$\underline{A} = \begin{bmatrix} \underline{a}_0' \\ \underline{A}_1' \\ \vdots \\ \underline{A}_p' \end{bmatrix} = \left.\begin{bmatrix} \underline{a}_1^0 & \cdots & \underline{a}_M^0 \\ \hline \underline{a}_{11}^1 & \cdots & \underline{a}_{M1}^1 \\ \vdots & & \vdots \\ \underline{a}_{1M}^1 & \cdots & \underline{a}_{MM}^1 \\ \hline \vdots & & \vdots \\ \hline \underline{a}_{11}^p & \cdots & \underline{a}_{M1}^p \\ \vdots & & \vdots \\ \underline{a}_{1M}^p & \cdots & \underline{a}_{MM}^p \end{bmatrix}\right\} R = pM + 1 \text{ rows}$$

Hence, $\underline{a}_i^0 = \underline{A}[1, i]$ and $\underline{a}_{ij}^k = \underline{A}[1 + (k-1)M + j, i]$ for $i, j = 1, \ldots, M$ and $k = 1, \ldots, p$.

Since $\underline{\alpha} = \text{vec}(\underline{A})$, any element $\underline{A}[m, n] = \underline{\alpha}[m + (n-1)R]$, where $R = pM + 1$.

Thus

$$\underline{a}_i^0 = \underline{A}[1, i] = \alpha[1 + (i-1)R]$$

and

$$\underline{a}_{ij}^k = \underline{A}[1 + (k-1)M + j, i] = \underline{\alpha}[1 + (k-1)M + j + (i-1)R].$$

Recall that all elements $\underline{a}_i^0$ and $\underline{a}_{ij}^k$ are set to zero except for the case $i = j$ and $k = 1$. Hence $\underline{\alpha}$ is a vector of zeros except for the elements

$$\underline{\alpha}[1 + i + (i-1)R] = \underline{A}[1 + i, i] = \underline{a}_{ii}^1 = \delta_i.$$

Implementation of the priors

How to determine $\underline{S}$

The prior mean for Σ implies

$$\mathrm{E}(\Sigma) = \frac{\underline{S}}{\underline{\nu} - M - 1} = \hat{\Sigma}_M \quad \Rightarrow \quad \underline{S} = (\underline{\nu} - M - 1)\hat{\Sigma}_M$$

Since $\hat{\Sigma}_M$ is a diagonal matrix with elements $\hat{\sigma}_{ii}$ on the main diagonal, $\underline{S}$ is also a diagonal matrix with non-zero elements only on the main diagonal

$$\underline{S}[i,i] = (\underline{\nu} - M - 1)\hat{\sigma}_{ii}$$

Implementation of the priors

How to determine $\underline{V}$

Recall that

$$\text{Var}(\underline{\alpha}) = \frac{\underline{S} \otimes \underline{V}}{\underline{\nu} - M - 1} = \frac{1}{\underline{\nu} - M - 1} \begin{bmatrix} \underline{S}[1,1]\underline{V} & & & \\ & \underline{S}[2,2]\underline{V} & & \\ & & \ddots & \\ & & & \underline{S}[M,M]\underline{V} \end{bmatrix}$$

which is block-diagonal with $(R \times R)$ dimensional blocks $\underline{S}[i,i]\underline{V}$ because the non-diagonal elements of $\underline{S}$ are zero. Using $\underline{S}[i,i] = (\underline{\nu} - M - 1)\hat{\sigma}_{ii}$ from above yields

$$\text{Var}(\underline{\alpha}) = \begin{bmatrix} \hat{\sigma}_{11}\underline{V} & & & \\ & \hat{\sigma}_{22}\underline{V} & & \\ & & \ddots & \\ & & & \hat{\sigma}_{MM}\underline{V} \end{bmatrix}$$

To make $\text{Var}(\underline{\alpha})$ diagonal (=all covariances are zero), we require $\underline{V}$ to be diagonal, too.

Note that $\underline{V}$ corresponds to the vector of all parameters of a single VAR equation and thus a column of A.

To pin down the elements in $\underline{V}$, we impose the variance assumptions and solve for $\underline{V}$.

The intercept corresponds to the first element in $\underline{V}$:

$$\text{Var}(a_i^0) = \text{Var}(\alpha[1 + (i-1)R]) = \hat{\sigma}_{ii}\underline{V}[1,1] \overset{!}{=} \theta_2^2\lambda^2\hat{\sigma}_{ii} \quad \Rightarrow \quad \underline{V}[1,1] = \lambda^2\theta_2^2$$

For the other parameters we have

$$\begin{aligned}\text{Var}(a_{ij}^k) &= \text{Var}(\alpha[1 + (k-1)M + j + (i-1)R]) \\ &= \hat{\sigma}_{ii}\underline{V}[1 + (k-1)M + j, 1 + (k-1)M + j] \overset{!}{=} \frac{\lambda^2}{k^2}\frac{\hat{\sigma}_{ii}}{\hat{\sigma}_{jj}}\end{aligned}$$

$$\Rightarrow \underline{V}[1 + (k-1)M + j, 1 + (k-1)M + j] = \frac{\lambda^2}{k^2\hat{\sigma}_{jj}}$$

Note that the elements of $\underline{V}$ do not depend on the equation i.

In matrix form this yields the diagonal matrix

$$\underline{V} = \lambda^2 \begin{bmatrix} \theta_2^2 & & & & & & & & & & & \\ & \frac{1}{\hat{\sigma}_{11}} & & & & & & & & & & \\ & & \ddots & & & & & & & & & \\ & & & \frac{1}{\hat{\sigma}_{MM}} & & & & & & & & \\ & & & & \frac{1}{2^2\hat{\sigma}_{11}} & & & & & & & \\ & & & & & \ddots & & & & & & \\ & & & & & & \frac{1}{2^2\hat{\sigma}_{MM}} & & & & & \\ & & & & & & & \ddots & & & & \\ & & & & & & & & \ddots & & & \\ & & & & & & & & & \frac{1}{p^2\hat{\sigma}_{11}} & & \\ & & & & & & & & & & \ddots & \\ & & & & & & & & & & & \frac{1}{p^2\hat{\sigma}_{MM}} \end{bmatrix}$$

Setting the dummy observations

To implement the estimation using the augmented data set, we need $T_0 = pM + M + 1$ dummy observations Y_0 and X_0 that satisfy

$$\underline{\alpha} = \text{vec}(\hat{A}_0) = \text{vec}[(X_0'X_0)^{-1}X_0'Y_0]$$
$$\underline{V} = (X_0'X_0)^{-1}$$
$$\underline{S} = (Y_0 - X_0\hat{A}_0)'(Y_0 - X_0\hat{A}_0)$$

BGR choose

$$Y_0 = \begin{bmatrix} 0_{1\times M} \\ \hline \frac{1}{\lambda}\text{diag}(\delta_1\hat{\sigma}_{11}^{\frac{1}{2}}, \ldots, \delta_M\hat{\sigma}_{MM}^{\frac{1}{2}}) \\ 0_{M(p-1)\times M} \\ \hline \text{diag}(\hat{\sigma}_{11}^{\frac{1}{2}}, \ldots, \hat{\sigma}_{MM}^{\frac{1}{2}}) \end{bmatrix}, \quad X_0 = \begin{bmatrix} (\lambda\theta_2)^{-1} & 0_{1\times pM} \\ \hline 0_{pM\times 1} & J_p \otimes \frac{1}{\lambda}\text{diag}(\hat{\sigma}_{11}^{\frac{1}{2}}, \ldots, \hat{\sigma}_{MM}^{\frac{1}{2}}) \\ \hline 0_{M\times 1} & 0_{M\times pM} \end{bmatrix},$$

where $J_p = \text{diag}(1, 2, \ldots, p)$. It is straightforward to show that these choices satisfy the above conditions.

Setting the shrinkage parameter λ

Recall that λ governs the overall amount of shrinkage:

- $\lambda = 0$: prior variances $= 0 \Rightarrow$ posterior $=$ prior (maximal shrinkage)
- $\lambda \to \infty$: prior variances $\to \infty \Rightarrow$ posterior $=$ ML estimator (no shrinkage)

BGR set λ data-dependent to prevent overfitting: the higher the number of variables, the smaller λ.
They proceed as follows:

- In a pre-evaluation period (1960:1 to 1969:12), they determine the average fit of the small model for EMPL, CPI, FFR.
- For each model, they choose λ such that it yields the same average fit for EMPL, CPI, FFR in the pre-evaluation period.

Forecast experiment

Setting $p = 13$, BGR conduct an out-of-sample forecast experiment with rolling 10-year estimation sample:

- Estimate the models from the sample 1960:1 to 1969:12 and compute 1, 3, 6, 12-month mean forecasts.
- Estimate the models from the sample 1960:2 to 1970:01 and compute 1, 3, 6, 12-month mean forecasts.
- Estimate the models from the sample 1960:3 to 1970:02 and compute 1, 3, 6, 12-month mean forecasts.
- Go on until last observation 2003:12.

Compute MSFE's for EMPL, CPI, FFR obtained from each model relative to a random walk with drift forecast.

Results: Does size matter?

Table I. BVAR, Relative MSFE, 1971–2003

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

Notes: The table reports MSFE relative to that from the benchmark model (random walk with drift) for employment (EMPL), CPI and federal funds rate (FFR) for different forecast horizons h and different models. *SMALL*, *CEE*, *MEDIUM* and *LARGE* refer to VARs with 3, 7, 20 and 131 variables, respectively. λ is the shrinkage hyperparameter.

Source: Banbura, Giannone and Reichlin (2010), Large Bayesian Vector Auto Regressions, Journal of Applied Econometrics 25(1), p. 78.

Results: BVAR or OLS with automatic lag length selection?

Table II. OLS and BVAR, relative MSFE, 1971–2003

		SMALL			*CEE*			*LARGE*
		$p = 13$	$p = \text{BIC}$	BVAR	$p = 13$	$p = \text{BIC}$	BVAR	BVAR
$h = 1$	EMPL	1.14	0.73	1.14	7.56	0.76	0.67	0.46
	CPI	0.89	0.55	0.89	5.61	0.55	0.52	0.50
	FFR	1.86	0.99	1.86	6.39	1.21	0.89	0.75
$h = 3$	EMPL	0.95	0.76	0.95	5.11	0.75	0.65	0.38
	CPI	0.66	0.49	0.66	4.52	0.45	0.41	0.40
	FFR	1.77	1.29	1.77	6.92	1.27	1.07	0.94
$h = 6$	EML	1.11	0.90	1.11	7.79	0.78	0.78	0.50
	CPI	0.64	0.51	0.64	4.80	0.44	0.41	0.40
	FFR	2.08	1.51	2.08	15.9	1.48	1.30	1.29
$h = 12$	EMPL	1.02	1.15	1.02	22.3	0.82	1.21	0.78
	CPI	0.83	0.56	0.83	21.0	0.53	0.57	0.44
	FFR	2.59	1.59	2.59	47.1	1.62	1.71	1.93

Notes: The table reports MSFE relative to that from the benchmark model (random walk with drift) for employment (EMPL), CPI and federal funds rate (FFR) for different forecast horizons h and different models. *SMALL*, *CEE* refer to the VARs with 3 and 7 variables, respectively. Those systems are estimated by OLS with number of lags fixed to 13 or chosen by the BIC. For comparison, the results of Bayesian estimation of the two models and of the large model are also provided.

Source: Banbura, Giannone and Reichlin (2010), Large Bayesian Vector Auto Regressions, Journal of Applied Econometrics 25(1), p. 79.