Bayesian Estimation of VAR and DSGE Models*

David Murakami†

11th June 2021

Contents

*Personal notes for the University of Oxford MPhil in Economics Advanced Macroeconomics II course. These notes are based on the lecture material by Bjorn van Roye, and supplemented by material from Joshua C.C. Chan, Silvia Miranda-Agrippino and Giovanni Ricco. Any typos or errors are my own.

†University of Oxford, St Cross College. Email: david.murakami@economics.ox.ac.uk.

1 Introduction

To quote Miao (2020): There are several different formal and informal econometric procedures to evaluate dynamic stochastic general equilibrium (DSGE) models quantitatively: Kydland and Prescott (1982) advocate a calibration procedure which was dominant in the early literature on Real Business Cycle (RBC) theory and analysis; Christiano and Eichenbaum (1992) use the generalised method of moments (GMM), pioneered by Hansen (1982), to estimate equilibrium relationships; Rotemberg and Woodford (1997) and Christiano, Eichenbaum, and Evans (2005) use the minimum distance estimation method based on the discrepancy between vector autoregression (VAR) and DSGE model impulse response functions (IRFs); Kim (2000) implemented full-information likelihood-based estimation methods; and Ireland (2004) described a hybrid method that combines DSGE models and VAR methods.

In these notes, we introduce the Bayesian estimation of DSGE models, with a brief introduction of Bayesian-VAR (BVAR) methods. This method has several advantages over other methods, as discussed in detail by Fernández-Villaverde (2010). First, unlike the GMM estimation based on equilibrium

relationships (such as the consumption Euler equation), Bayesian analysis is system based and fits the solved DSGE model to a vector of aggregate time series. Second, the estimation is based on the likelihood function generated by the DSGE model rather than, for instance, the discrepancy between DSGE model responses and VAR IRFs. Third, maximising the likelihood function with the ML method is challenging. By contrast, computing posteriors with Bayesian methods is much easier. Fourth, prior distributions can be used to incorporate additional information into the parameter estimation.

Our focus here will be on linearised DSGE models. Good readings for Bayesian estimation of both linearised and nonlinear models are An and Schorfheide (2007), DeJong and Dave (2012), and Fernández-Villaverde (2010).

2 Principles of Bayesian Estimation

2.1 Bayesian vs frequentist approach

But first, a recap: we're interested in estimating the parameter θ, which can be thought of as either a scalar or a vector. The frequentist approach – which is what most, if not all, graduate students of economics are familiar with, seeks to estimate θ from data, say, y, from the population sampled. y contains all the information we know about θ, and it is used to form the likelihood function. This data information is used to obtain an estimate of θ, $\hat{\theta}$, and it's typically done so with the method of ML. The estimate, $\hat{\theta}$, is referred to as the ML estimator (MLE), as it maximises the value of the likelihood function.

The Bayesian approach is to form the prior belief that θ is unknown. That is, we have fundamental uncertainty about it, and that there is no single true value for θ. So, Bayesians treat θ as a random variable, and assign it its own a priori probability distribution. This prior distribution is then combined with the distribution from the data (and hence, the likelihood function) to form what Bayesian statisticians call the "posterior distribution." This posterior distribution is of primary interest to the Bayesian statistician, although macroeconomists are more so interested in the means of these posterior distributions when it comes to estimating model parameters of a DSGE model – e.g., see Smets and Wouters (2003, 2007) and Christiano, Trabandt, et al. (2011).

Thus, the Bayesian approach can be summarised as:

1. Sample data from the population.
2. Compute the likelihood function.
3. Observe data results from realisations of θ from its probability distribution.
4. Form a prior belief of what the distribution of θ may look like – i.e., specify a prior distribution (mean, variance, and the distribution's "hyperparameters").

5. Estimate the posterior distribution using Bayes' Theorem, by combining the likelihood and the prior distribution.

Table 1: Bayesian vs Frequentist Approach

	Frequentist	Bayesian
Parameter θ	Unique true value	Random variable
Object of interest	θ	posterior distribution of θ
Information	Sample data	Sample data and prior distribution
Estimation	ML	Bayes' Rule + ML
Estimate of θ	$\hat{\theta}$	Posterior distribution of θ

2.2 Bayes' Theorem

Here, we derive Bayes' Theorem or Bayes' Rule, an idea which we will use time and time again.

Recall that for two events, A and B, the probability of A occurring, conditional on B, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

where $A \cap B$ represents the intersection of events A and B.

Consider the following example of a dice roll. The set of possible outcomes is denoted by $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let A be "a number > 3", so $P(A) = (\{4, 5, 6\}) = \frac{1}{2}$. Let B be "an even number", so $P(B) = (\{2, 4, 6\}) = \frac{1}{2}$. Thus, $P(A \cap B) = P(\{4, 6\}) = \frac{1}{3}$. The conditional probability is then $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{2}{3}$.

So, doing a little algebra we have

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A \cap B)}{P(B)} \frac{P(A)}{P(A)} \\ &= \frac{P(A \cap B)}{P(A)} \frac{P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B)}, \end{aligned} \tag{1}$$

which is Bayes' Theorem for events. Bayes' Theorem handles conditional probabilities by connecting $P(A|B)$ and $P(B|A)$: Initially there is an unconditional probability, $P(A)$. But then event B occurs, and gives additional information, $P(B|A)$, allowing us to update the belief that A will occur. Bayes' rule essentially updates the belief $P(A|B)$.

Consider the dice rolling example again. We are interested in determining $P(A|B)$, the probability

of a number larger than 3 is rolled, given that we roll an even number:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\frac{2}{3}\frac{1}{2}}{\frac{1}{2}} = \frac{2}{3}.$$

Recall that the unconditional probability of a number larger than 3 is $P(A) = \frac{1}{2}$. So here our initial belief has been updated with information from B (the rolled number being even).

Now, consider probability densities. Suppose the probability density function (PDF) of X is denoted by $f(x)$, and the PDF of Z is $f(z)$, with joint PDF, $f(x, z)$. The conditional density of X is then

$$f(x|z) = \frac{f(x, z)}{f(z)}.$$

Now, replace x, z with the random variables θ, y (our parameter and data) by setting $x = \theta$ and $z = y$. We then get Bayes' Rules for densities:

$$\begin{aligned} f(\theta|y) &= \frac{f(\theta, y)}{f(y)} \\ &= \frac{f(y|\theta)f(\theta)}{f(y)}, \end{aligned} \tag{2}$$

and note that $f(\theta|y)$ is the posterior distribution, $f(y|\theta)$ is the likelihood function, $f(\theta)$ is the prior distribution, and $f^{-1}(y)$ is the marginal likelihood. We have some more to say about this, but first let's move onto the vector case.

2.2.1 The vector case

Now, consider the vector case (which will be the default going forward). Suppose we want to estimate the parameter vector $\boldsymbol{\theta} \in \Theta$. Let the prior density be $f(\boldsymbol{\theta})$, and suppose we have data $\mathbf{y} = \{y_1, y_2, ..., y_T\}$. The likelihood density is the density of data $\mathbf{y}$ given by $\boldsymbol{\theta}$:

$$f(\mathbf{y}|\boldsymbol{\theta}) = f(y_1|\boldsymbol{\theta}) \prod_{t=2}^{T} f(y_t|\mathbf{y}_{-1}, \boldsymbol{\theta}).$$

The likelihood function of the parameter is given by

$$L(\mathbf{y}; \boldsymbol{\theta}) \equiv f(\mathbf{y}|\boldsymbol{\theta}).$$

The ML method of estimation is to search for a parameter vector that will maximise $L(\mathbf{y}; \boldsymbol{\theta})$.

Using Bayes' Theorem, we can compute the posterior as

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y},\boldsymbol{\theta})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{3}$$

But the marginal likelihood (and indeed the likelihood function itself) may be complex or impossible to analytically solve. We proceed as follows.

Define the **posterior kernel** as $\mathcal{K}(\boldsymbol{\theta}|\mathbf{y})$, where

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \equiv \mathcal{K}(\boldsymbol{\theta}|\mathbf{y}). \tag{4}$$

We can do this because $f(\mathbf{y})$ (or, equivalently, $f^{-1}(\boldsymbol{\theta})$) can be treated as a proportional constant – i.e., it does not contain $\boldsymbol{\theta}$, and so it does not change as $\boldsymbol{\theta}$ changes. Thus, the Bayesian approach is to choose a parameter vector $\boldsymbol{\theta}$ so as to maximise the posterior density, $f(\boldsymbol{\theta}|\mathbf{y})$, or, equivalently, the posterior kernel, $\mathcal{K}(\boldsymbol{\theta}|\mathbf{y})$.

As mentioned, the difficulty of using the ML method or the Bayesian method is that the likelihood function typically has no analytical solution. Also, Bayesian analysis involves the computation of the conditional distribution of a function of the parameters, $h(\boldsymbol{\theta})$:

$$\mathbb{E}\left[h(\boldsymbol{\theta})|\mathbf{y}\right] = \int h(\boldsymbol{\theta})f(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}.$$

Numerical integration is therefore needed. To implement the Bayesian method, we use a filtering procedure to evaluate the likelihood function. We then simulate the posterior kernel using a Markov chain Monte Carlo (MCMC) method such as the Metropolis-Hastings algorithm or Gipps sampler (which we shall discuss later).

Figure 1: The Posterior Distribution

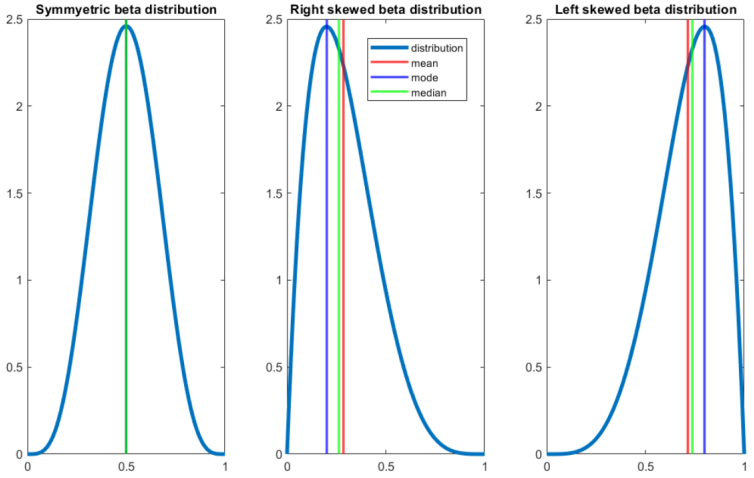
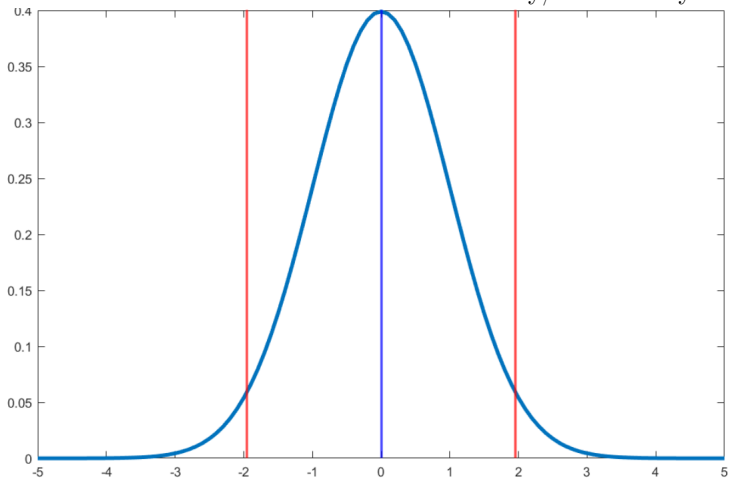


Figure 2: The Posterior Distribution – Credibility/Uncertainty Bands



2.2.2 Example: Binomial distribution

Let's look at a simple example. Suppose we flip a coin n times, and observe m successes ("heads") over the n flips. Every flip can result in either "heads" or "tails", and we are interested in the probability of success, $\theta = p$. As we know, the binomial distribution is given by

$$f(y_i|p) = \binom{n}{m} p^{y_i}(1-p)^{1-y_i}.$$

Assuming that we have independent coin flips (and a fair coin), the density for the n flips together is just the product of the individual densities of each of the n flips:

$$\begin{aligned} f(y_1, ..., y_n|p) &= \prod_{i=1}^{n} f(y_i|p) \\ \Leftrightarrow f(\mathbf{y}|p) = L(\mathbf{y}|p) &= \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}, \end{aligned} \tag{5}$$

which represents the likelihood function of the data. The log-likelihood is simply the log of the likelihood function:

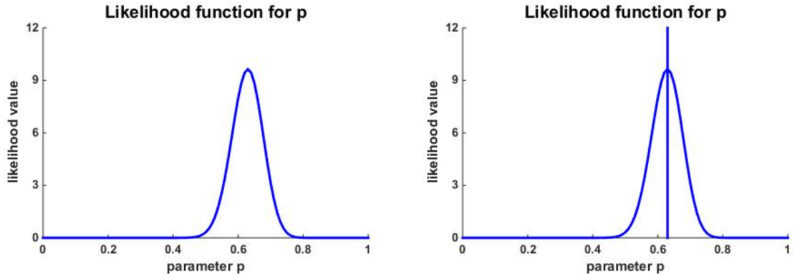
$$\log L(\mathbf{y}|p) = l(\mathbf{y}|p) = \sum_{i=1}^{n} y_i \ln p + \left(n - \sum_{i=1}^{n} y_i\right) \ln(1-p),$$

and if we differentiate the log-likelihood function we attain a first-order condition (FOC):

$$\begin{aligned} \frac{\partial l(\mathbf{y}|p)}{\partial p} &= \frac{1}{p}\sum_{i=1}^{n} y_i + \frac{1}{1-p}\left(n - \sum_{i=1}^{n} y_i\right) = 0 \\ \implies 0 &= (1-\hat{p})\sum_{i=1}^{n} y_i + p\left(n - \sum_{i=1}^{n} y_i\right) \\ \hat{p} &= \frac{\sum_{i=1}^{n} y_i}{n} = \frac{m}{n}. \end{aligned}$$

Now, suppose had $n = 100$, and we observed $m = 63$ "heads". Then $\hat{p} = 0.63$ (which also coincides with the sample mean, $\bar{y}$).

Figure 3: Outcome of an Experiment



So how can we improve this using Bayesian methods? Well, for starters, if we believe that the coin is unbiased, then we may have a prior belief that p should be clustered around 0.5. Furthermore, as it's a probability, the support of this prior distribution should be $[0, 1]$. A candidate prior distribution is the **Beta distribution** which has support $[0, 1]$ and is flexible in terms of its shape.

Definition 1 (**Beta Distribution**)**.** The density function of the Beta distribution is

$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x^{\beta-1}),$$

where $\Gamma(\cdot)$ is the Gamma function, and α and β are hyperparameters which determine the shape of the density. The Beta distribution has the following mean,

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta},$$

and variance,

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

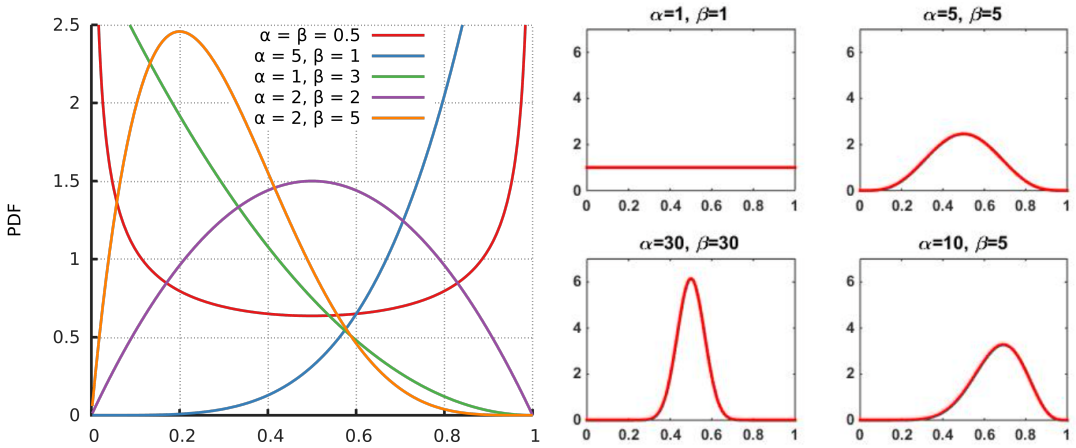
and moment generating function (MGF),

$$M_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}.$$

As the hyperparameters α and β vary, the Beta distribution takes on many shapes, as shown in Figure 4. The PDF can be strictly increasing ($\alpha > 1, \beta = 1$), strictly decreasing ($\alpha = 1, \beta > 1$), U-shaped ($\alpha < 1, \beta < 1$), or unimodal ($\alpha > 1, \beta > 1$). The case $\alpha = \beta$ yields a symmetric PDF about 0.5 with mean 0.5 and variance $(4(2\alpha+1))^{-1}$, and when $\alpha = \beta = 1$, the Beta distribution reduces down to uniform distribution, $U(0,1)$.

One thing to note: there is no hard and fast rule (that I know of) to select hyperparameters. Their selection is made to simply reflect personal beliefs about the distribution. For our coin flip example, a reasonable prior for a coin is that it should not be more or less fair, and that if it is biased, the bias should not be too large. So, the prior should be centred around 0.5, with a small variance, which implies that we set $\alpha = \beta$ with high values (to get a "tight" prior). Here, $\alpha = \beta = 40$ seems to be acceptable.

Figure 4: Density of the Beta Distribution



For our coin flip experiment, the density of Beta distribution can be written as

$$f(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$
$$\implies f(p|\alpha,\beta) \propto p^{\alpha-1}(1-p)^{\beta-1}, \tag{6}$$

where we drop the quotient of Gamma functions as it does not contain any information pertaining to the parameter of interest, p. Now, apply Bayes' Rule:

$$f(p|\mathbf{y}) = \frac{f(\mathbf{y}|p)f(p)}{\int f(\mathbf{y}|p)f(p)dp},$$

and then focus on the posterior kernel,

$$\begin{aligned} f(p|\mathbf{y}) &\propto f(\mathbf{y}|p)f(p) \equiv \mathcal{K}(p|\mathbf{y}) \\ &\propto p^m(1-p)^{n-m} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{m+\alpha-1}(1-p)^{n-m+\beta-1}. \end{aligned} \tag{7}$$

Wrap up by defining $\bar{\alpha} = m+\alpha$ and $\bar{\beta} = n-m+\beta$, and writing the posterior distribution (proportionally as the kernel) as

$$f(p|\mathbf{y}) \propto p^{\bar{\alpha}-1}(1-p)^{\bar{\beta}-1}. \tag{8}$$

Below is some sample R code which works through this example:

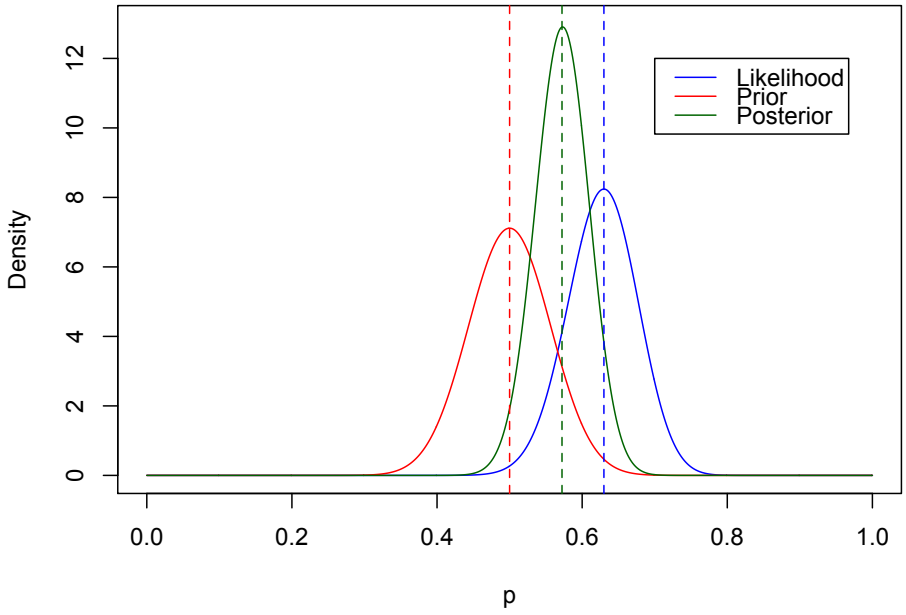
```
rm(list=ls(all=TRUE));
graphics.off();

p.seq = seq(0,1,by=0.001);
# Plot likelihood
plot(p.seq,100*dbinom(x=63,prob=p.seq,size=100),type="l",col="blue",
xlab="p",ylab="Density",ylim=c(0,13)) #scale factor of 100 for plotting purposes
# Find MLE
Lik = function(p) prod((dbinom(63,100,p,log=FALSE)));
optimise(Lik,lower=0,upper=1,maximum=TRUE)

## $maximum
## [1] 0.6299815
##
## $objective
## [1] 0.08240399

# Update prior
prior = dbeta(x=p.seq,40,40);
lines(p.seq,prior,col="red");
# Plot posterior
posterior = 100*dbinom(x=63,prob=p.seq,size=100)*prior;
lines(p.seq,posterior,col="darkgreen");
# Add mean lines
abline(v=optimise(Lik,lower=0,upper=1,maximum=TRUE)$maximum,col="blue",lty=2);
abline(v=1/2,col="red",lty=2) #alpha/(alpha+beta)
abline(v=103/(103+77),col="darkgreen",lty=2) #bar.alpha/(bar.alpha+bar.beta)
# Add legend
legend(0.7,12,legend=c("Likelihood","Prior","Posterior"),
col=c("blue","red","darkgreen"),lty=1);
```

Figure 5: Likelihood, Prior, Posterior Distribution for Binomial Example



2.2.3 Example: Normal prior and a simple white noise process

Here's another simple example to see that Bayesian estimation essentially lies between calibration and maximum likelihood estimation. Suppose that a DGP is given by

$$y_t = \mu + \epsilon_t,$$

where $\epsilon_t \sim \mathcal{N}(0,1)$ is a Gaussian white noise process. We want to estimate the parameter μ from sample data $\mathbf{y}$. We can compute the likelihood density as:

$$p(\mathbf{y}|\mu) = (2\pi)^{-T/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}(y_t - \mu)^2\right).$$

We then obtain the ML estimate

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t,$$

which is nothing but the sample average.

Suppose that the prior is Gaussian with mean μ_0 and variance σ_μ^2. Then the posterior kernel is

given by

$$\left(2\pi\sigma_{\mu}^{2}\right)^{-1/2}\exp\left(-\frac{1}{2\sigma_{\mu}^{2}}(\mu-\mu_{0})^{2}\right)(2\pi)^{-T/2}\exp\left(-\frac{1}{2}\sum_{t=1}^{T}(y_{t}-\mu)^{2}\right).$$

Thus, the Bayesian estimate is given by

$$\hat{\mu}_{BE}=\frac{T\hat{\mu}+\sigma_{\mu}^{-2}\mu}{T+\sigma_{\mu}^{-2}},$$

which is a linear combination of the prior mean and the MLE. When we have no prior information (i.e., $\sigma_{\mu}^{2}\to\infty$), the Bayesian estimate converges to the ML estimate. When we are sure that the prior calibrated value is true (i.e., $\sigma_{\mu}^{2}\to 0$), then $\hat{\mu}_{BE}\to\mu_{0}$.

3 Bayesian Estimation of Linear Regressions

We've covered the simple base premise of Bayesian estimation: Use Bayes' Theorem to get a posterior distribution of an object of interest by finding the product of its likelihood function and its prior distribution. In this section, we extend our analysis to some simple linear regression models and show that we follow the aforementioned steps.

3.1 The simple linear regression model

Let's consider a classic, simple linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$

where I use the notation from Davidson and MacKinnon (2004), so $\mathbf{y}$ is a $n \times 1$ vector, $\mathbf{X}$ is an $n \times k$ matrix of regressors, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients, and $\mathbf{u}$ is the vector of NID errors, so each i-th row looks like $y_i = \mathbf{X}_i\boldsymbol{\beta} + u_i$.

Here, our parameters of interest are, of course, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. Using OLS, our estimates for these are the familiar expressions:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \\ \hat{\sigma}^2 &= \frac{\hat{\mathbf{u}}^\top\hat{\mathbf{u}}}{n-k}.\end{aligned}$$

To form our Bayesian estimate, let's first make a huge assumption (later on, we will relax this): σ^2 is a known constant. Then we next define the density of a multivariate normal distribution, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, as

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\},$$

and we know that since $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, it follows that $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Hence, we can write

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\theta}) &= (2\pi)^{-n/2}|\sigma^2\mathbf{I}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \\ f(\mathbf{y}|\boldsymbol{\beta}) &= \underbrace{(2\pi)^{-n/2}|\sigma^2\mathbf{I}|^{-1/2}}_{\text{constant}}\underbrace{\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}}_{\text{kernel}} \\ \implies f(\mathbf{y}|\boldsymbol{\beta}) &\propto \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}. \end{aligned} \tag{9}$$

Why do we do this? Well, σ^2 is a known scalar constant, so it's easier to just deal with the kernel

of the normal distribution. It's worth mentioning that so far we aren't doing anything new; we're just writing out an expression for the likelihood function of a normally distributed quantity. With the likelihood function in hand, we then need a prior distribution to compute the posterior distribution.

When considering what prior distribution to pick for $\boldsymbol{\beta}$, we need to look for a multivariate distribution as $\boldsymbol{\beta}$ is a vector. A natural choice is the multivariate normal distribution – and it's not too outlandish given our knowledge of both finite sample and asymptotic theory of the OLS estimator. So, let's assume that $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0)$, where $\boldsymbol{\beta}_0$ and $\boldsymbol{\Omega}_0$ are the prior mean vector and prior variance covariance matrix, respectively. Then the prior density of $\boldsymbol{\beta}$ can be written as

$$\begin{aligned} f(\boldsymbol{\beta}) &= (2\pi)^{-n/2}|\Omega_0|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\} \\ \implies f(\boldsymbol{\beta}) &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}. \end{aligned} \tag{10}$$

So now we can use Bayes' Theorem and substitute in our expressions, (9) and (10), to write

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\beta})f(\boldsymbol{\beta}) \\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right]\right\}. \end{aligned} \tag{11}$$

This is looking an absolute mess, but bare with it for a bit. Let's manipulate the term inside the exponent of the posterior,

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0),$$

and expand it out by "completing the squares":[1]

$$\begin{aligned} &\mathbf{y}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} - \mathbf{y}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X}\boldsymbol{\beta} - (\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} + (\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X}\boldsymbol{\beta} \\ &\quad + \boldsymbol{\beta}^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 \\ &= \mathbf{y}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} + (\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 \\ &= \mathbf{y}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} - 2\boldsymbol{\beta}^\top\left(\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} + \boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0\right) + \boldsymbol{\beta}^\top\left(\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X} + \boldsymbol{\Omega}_0^{-1}\right)\boldsymbol{\beta} + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0. \end{aligned}$$

[1] Recall that in the scalar case, this is

$$ax^2 + bx + c = a(x-h)^2 + k,$$

where $h = -\frac{b}{2a}$ and $k = c - ah^2$. In the matrix case, we have:

$$\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{B} + \mathbf{C} = (\mathbf{x}-\mathbf{H})^\top\mathbf{A}(\mathbf{x}-\mathbf{H}) + \mathbf{K},$$

where $\mathbf{H} = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{B}$ and $\mathbf{K} = \mathbf{C} - \frac{1}{4}\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}$, and $\mathbf{A}$ is symmetric.

Next, let's define

$$\bar{\boldsymbol{\beta}} = \bar{\boldsymbol{\Omega}} \left(\boldsymbol{\Omega}_0^{-1} \boldsymbol{\beta}_0 + \mathbf{X}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{y} \right), \tag{12}$$

$$\bar{\boldsymbol{\Omega}} = \left(\boldsymbol{\Omega}_0^{-1} + \mathbf{X}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{X} \right)^{-1}, \tag{13}$$

and let's do a little "add and subtract":

$$\begin{aligned} &\mathbf{y}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{y} - 2\boldsymbol{\beta}^\top \underbrace{\bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\Omega}}}_{\mathbf{I}} \left(\mathbf{X}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{y} + \boldsymbol{\Omega}_0^{-1} \boldsymbol{\beta}_0 \right) \\ &+ \boldsymbol{\beta}^\top \left(\mathbf{X}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{X} + \boldsymbol{\Omega}_0^{-1} \right) \boldsymbol{\beta} + \boldsymbol{\beta}_0^\top \boldsymbol{\Omega}_0^{-1} \boldsymbol{\beta}_0 \underbrace{+ \bar{\boldsymbol{\beta}}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}}}_{=0}, \end{aligned}$$

and so we can clean up our disgusting expression a bit:

$$\begin{aligned} &\mathbf{y}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{y} - 2\boldsymbol{\beta}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}} + \boldsymbol{\beta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\beta} + \boldsymbol{\beta}_0^\top \boldsymbol{\Omega}_0^{-1} \boldsymbol{\beta}_0 + \bar{\boldsymbol{\beta}}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}} \\ &= \left[\boldsymbol{\beta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\beta} + \bar{\boldsymbol{\beta}}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}} - 2\boldsymbol{\beta}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}} \right] + \mathbf{y}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{y} + \boldsymbol{\beta}_0^\top \boldsymbol{\Omega}_0^{-1} \boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}} \\ &= (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top \bar{\boldsymbol{\Omega}}^{-1} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + \mathbf{y}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{y} + \boldsymbol{\beta}_0^\top \boldsymbol{\Omega}_0^{-1} \boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}}. \end{aligned}$$

So now we can express (11) as

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{y}) &\propto \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top \bar{\boldsymbol{\Omega}}^{-1} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + \mathbf{y}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{y} + \boldsymbol{\beta}_0^\top \boldsymbol{\Omega}_0^{-1} \boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}} \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top \bar{\boldsymbol{\Omega}}^{-1} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \right] \right\} \exp\left\{ -\frac{1}{2} \left[\mathbf{y}^\top (\sigma^2 \mathbf{I})^{-1} \mathbf{y} + \boldsymbol{\beta}_0^\top \boldsymbol{\Omega}_0^{-1} \boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}^\top \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\beta}} \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top \bar{\boldsymbol{\Omega}}^{-1} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \right] \right\}, \end{aligned} \tag{14}$$

and we're done! We once again ignore constants as we're only interested in the distribution of $\boldsymbol{\beta}$. Thus, we have managed to calculate our posterior distribution, along with getting an expression for its mean and variance. It looked extremely messy, but we essentially followed the procedure outlined in the previous section. We first attained a likelihood function for the sample data, $f(\mathbf{y}|\boldsymbol{\beta})$ (remember that we first had $f(\mathbf{y}|\boldsymbol{\theta})$ but we ignored constant elements and focused on the kernel), then we determined a prior distribution for our object of interest, $f(\boldsymbol{\beta})$, and then we merged the two elements together to form our posterior distribution, $f(\boldsymbol{\beta}|\mathbf{y})$.

There are a couple things worth mentioning at this stage: notice that in this simple example, the expression in the exponent of (14) is a quadratic in $\boldsymbol{\beta}$. Therefore, $\boldsymbol{\beta}|\mathbf{y}$ is normally distributed (Gaussian). In fact, we can express the distribution of this as stuff that we know: $\boldsymbol{\beta}|\mathbf{y} \sim \mathcal{N}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\Omega}})$. When the posterior and prior share the same class of distribution, the prior is referred to as a **conjugate prior**.

Definition 2 (**Conjugate Distributions**)**.** In Bayesian probability theory, if the posterior distribution, $f(\boldsymbol{\theta}|\mathbf{y})$, are in the same probability distribution family as the prior probability distribution, $f(\boldsymbol{\theta})$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function, $f(\mathbf{y}|\boldsymbol{\theta})$.

3.1.1 Example: Estimating the Taylor Rule

Consider Taylor's (1993) original interest rate rule:

$$r_t = \bar{\pi}_t + \frac{1}{2}(\bar{\pi}_t - 2) + \frac{1}{2}(y_t - y_t^*) + 2,$$

where r_t is the Federal Funds Rate (FFR), $\bar{\pi}_t$ is annual inflation, and y_t^* is trend (log) GDP. After doing a bit of rearranging and defining x_t as the (real) output gap, we can rewrite the Taylor Rule as

$$r_t = 1 + 1.5\bar{\pi}_t + \frac{1}{2}x_t, \tag{15}$$

which is now a linear function which we can estimate using regression. Using data from FRED between 1980Q1 to 2006Q4, and using the HP filter, we regress the FFR on inflation and the output gap to get the following regression results:

$$r_t = \underset{(1.20,2.79)}{2.00} + \underset{(0.97,1.31)}{1.14}\,\bar{\pi}_t + \underset{(-0.03,0.51)}{0.24}\,x_t + \hat{\epsilon}_t, \tag{16}$$

where, clearly, we have

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1.14 \\ 0.24 \end{bmatrix}.$$

Now, what if we want to use Bayesian estimation? We just attained our estimates using OLS, so next we need to construct a distribution for our priors.

For starters, let's suppose that we use Taylor's original proposed rule, (15), as our prior where we believe the following:

$$\boldsymbol{\beta}_0 = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1.5 \\ 0.5 \end{bmatrix}.$$

Just as a reminder: β_1 is the target of the Fed, β_2 is the coefficient on inflation, and β_3 is the coefficient

on the output gap. Next, for our prior variance we have the following 3×3 matrix, $\mathbf{\Omega}_0$:

$$\mathbf{\Omega}_0 = \begin{bmatrix} \omega_1^2 & 0 & 0 \\ 0 & \omega_2^2 & 0 \\ 0 & 0 & \omega_3^2 \end{bmatrix},$$

where the diagonal terms are our hyperparameters for the variance terms of $\boldsymbol{\beta}_0$. We obviously have some uncertainty surrounding our prior $\boldsymbol{\beta}_0$: so let's suppose that assume values between 0.5 and 1.5 for β_1; 1.2 and 1.8 for β_2; and 0.3 and 0.07 for β_3. These values will of course give us standard deviations of 0.5, 0.3, and 0.2 for ω_1, ω_2, and ω_3, respectively:

$$\mathbf{\Omega}_0 = \begin{bmatrix} \omega_1^2 & 0 & 0 \\ 0 & \omega_2^2 & 0 \\ 0 & 0 & \omega_3^2 \end{bmatrix} = \begin{bmatrix} 0.25 & 0 & 0 \\ 0 & 0.09 & 0 \\ 0 & 0 & 0.04 \end{bmatrix}$$

Using these priors, our Bayesian estimation gives us

$$r_t = \underset{(0.92,2.13)}{1.52} + \underset{(1.10,1.38)}{1.23}\,\bar{\pi}_t + \underset{(0.13,0.57)}{0.35}\,x_t + \hat{\epsilon}_t, \tag{17}$$

where our vector Bayesian estimates for the parameter vector $\boldsymbol{\beta}$ is

$$\bar{\boldsymbol{\beta}} = \begin{bmatrix} 1.52 \\ 1.23 \\ 0.35 \end{bmatrix},$$

which are the median values of the posterior distributions. In Bayesian estimation, the intervals denoted in (17) are referred to as **uncertainty bands**, rather than our usual notation of "confidence intervals" – it's a minor point, but something worth noting. Also, here, in case it wasn't obvious, the uncertainty bands are of 95% uncertainty of a posterior normal distribution.

Comparing the Bayesian estimate, (17), with the our ML/OLS estimate, (16), and our prior (the original Taylor rule), (15), we note:

- The point estimate of our Bayesian estimation is the median of the posterior distribution.
- The prior has pushed the coefficients [of our ML/OLS estimates] closer to the original Taylor coefficients.
- The Bayesian estimate is a weighted average between the ML/OLS estimate and our prior.

Some alternative priors which we could've used are a **non-informative** (flat/loose/diffuse) prior:

$$\begin{aligned} &\beta_1 = \beta_2 = \beta_3 = 0, \\ \implies &\omega_1^2 = \omega_2^2 = \omega_3^2 \to \infty, \end{aligned}$$

or a **non-reasonable** prior:

$$\begin{aligned} &\beta_1 = \beta_2 = \beta_3 = 0, \\ &\omega_1^2 = \omega_2^2 = \omega_3^2 = 0.0025. \end{aligned}$$

These prior specifications would give the results as shown in Table 2. Notice how the diffuse prior produces ML estimates, while the non-reasonable prior ("bad prior") pushes all coefficients towards 0.

Table 2: Alternative Priors

Prior	β_1	β_2	β_3
Taylor Rule	1	1.5	0.5
ML/OLS	2.00 (1.20,2.79)	1.14 (0.97,1.31)	0.24 (−0.03,0.51)
"Good prior"	1.52 (0.92,2.13)	1.23 (1.10,1.38)	0.35 (0.13,0.57)
"Diffuse prior"	2.00 (1.20,2.79)	1.14 (0.97,1.31)	0.24 (−0.03,0.51)
"Bad prior"	0.17 (0.07,0.27)	0.77 (0.70,0.84)	0.01 (−0.09,0.10)

3.2 The regression model with unknown variance

Up until now we have made the very strong assumption that the error variance, σ^2, is known. Even in classical econometrics, this is an unrealistic assumption. If we assume that the error variance is unknown, as is usually the case, then our vector of parameters to be estimated is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. From (4), we can rewrite Bayes' Rule, or the posterior kernel, as

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}, \sigma^2), \tag{18}$$

where $f(\boldsymbol{\beta}, \sigma^2)$ is a joint prior. We can make the simplifying assumption of **prior independence**:

$$f(\boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{\beta}) f(\sigma^2), \tag{19}$$

where we can propose a separate prior for each parameter.

The posterior density, $f(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$, is trickier to handle. It's a joint posterior, which is not tractable as we need to estimate and disentangle the densities of $\boldsymbol{\beta}$ and σ^2 (in this example). Thus, we need the

marginal posterior for each parameter that we wish to estimate. This is simple in theory:

$$f(\boldsymbol{\beta}|\mathbf{y}) = \int_{-\infty}^{\infty} f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) d\sigma^2,$$
$$f(\sigma^2|\mathbf{y}) = \int_{-\infty}^{\infty} f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) d\boldsymbol{\beta},$$

but difficult in practice. Just consider the example we had in the previous section (but this time we include σ^2 in $\boldsymbol{\theta}$), so our likelihood and kernel densities are:

$$f(\mathbf{y}|\boldsymbol{\theta}) = (2\pi)^{-n/2}|\sigma^2\mathbf{I}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$
$$\implies f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) \propto |\sigma^2\mathbf{I}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}, \tag{20}$$

and our prior distribution for $\boldsymbol{\beta}$ was

$$f(\boldsymbol{\beta}) = (2\pi)^{-n/2}|\Omega_0|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\}$$
$$\implies f(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\}. \tag{21}$$

Now, we need a candidate prior distribution for σ^2. Since σ^2 is a variance term, we need a distribution with a strictly positive support. One candidate is the Inverse-Gamma distribution.

Definition 3 (**Inverse-Gamma Distribution**)**.** The density function of a random variable x which follows the Inverse-Gamma distribution, $X \sim \Gamma^{-1}(\alpha, \beta)$, is defined over the support $x > 0$ as

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha+1} \exp\left\{-\frac{\beta}{x}\right\},$$

where $\Gamma(\cdot)$ is the Gamma function, and α and β are hyperparameters which determine the shape and scale of the density, respectively. The Inverse-Gamma distribution has the following mean,

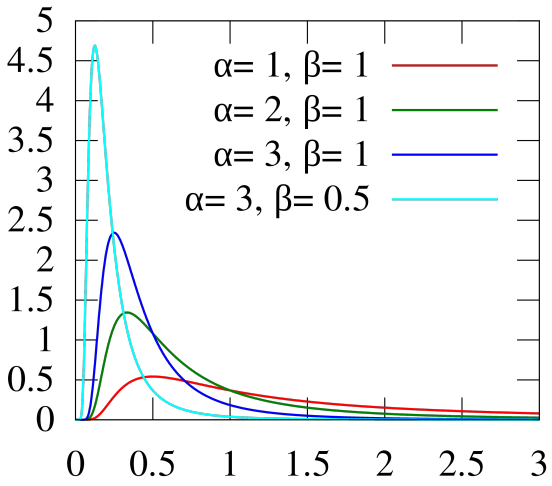
$$\mathbb{E}[X] = \frac{\beta}{\alpha - 1},$$

and variance,

$$\text{Var}(X) = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)}.$$

The MGF of the Inverse-Gamma distribution does not exist.

Figure 6: Inverse-Gamma PDF



Our prior distribution for σ^2 is

$$f(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma_0^2)^{-\alpha-1}\exp\left\{-\frac{\beta}{\sigma_0^2}\right\}$$
$$\implies f(\sigma^2) \propto (\sigma_0^2)^{-\alpha-1}\exp\left\{-\frac{\beta}{\sigma_0^2}\right\}. \tag{22}$$

Putting (20), (21), and (22) into (18), we get

$$f(\boldsymbol{\beta},\sigma^2|\mathbf{y}) \propto |\sigma^2\mathbf{I}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$
$$\times\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}(\sigma_0^2)^{-\alpha-1}\exp\left\{-\frac{\beta}{\sigma_0^2}\right\}. \tag{23}$$

The issue is that integration and obtaining an analytical solution to this problem is close to, if not, impossible. We will have to approach this problem numerically.

3.3 Markov Chain Monte Carlo methods

Recall that a Markov chain is a stochastic model that describes a sequence of events in which the probability of each event depends on the state of the previous event, while a Monte Carlo simulation is a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. Markov chain Monte Carlo (MCMC) algorithms, such as the Metropolis Hastings (MH) algorithm and the Gibbs sampler, have become extremely popular in statistics and econometrics as a

way of approximately sampling from complicated probability distributions in high dimensions. The MH algorithm, in particular, has become a standard process of estimating DSGE models.

3.3.1 The Gibbs sampler

Here, we narrow our focus on the Gibbs sampler. Assume for instance that you have a posterior distribution, $f(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$, for parameters $\boldsymbol{\beta}$ and σ^2, but that it is impossible to attain an analytical solution to the unconditional posterior distributions, $f(\boldsymbol{\beta}|\mathbf{y})$ and $f(\sigma^2|\mathbf{y})$. However, assume that it is possible to evaluate the conditional posterior distributions, $f(\boldsymbol{\beta}|\mathbf{y}, \sigma^2)$ and $f(\sigma^2|\mathbf{y}, \boldsymbol{\beta})$, and that these conditional posteriors correspond to known distributions. In such a case, we can conduct what is known as the **Gibbs algorithm**:

1. Fix starting values $\boldsymbol{\beta}_{(0)}$ and $\sigma^2_{(0)}$ for two parameters, $\boldsymbol{\beta}$ and σ^2.
2. Draw first value of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_{(1)}$, from the conditional posterior, $f(\boldsymbol{\beta}|\mathbf{y}, \sigma^2_{(0)})$.
3. Draw first value of σ^2, $\sigma^2_{(1)}$ from the conditional posterior, $f(\sigma^2|\mathbf{y}, \boldsymbol{\beta}_{(1)})$, using $\boldsymbol{\beta}_{(1)}$.
4. Start a new cycle: draw value $\boldsymbol{\beta}_{(2)}$ from the conditional posterior, $f(\boldsymbol{\beta}|\mathbf{y}, \sigma^2_{(1)})$, using $\sigma^2_{(1)}$.
5. Draw value $\sigma^2_{(2)}$ from the conditional posterior, $f(\sigma^2|\mathbf{y}, \boldsymbol{\beta}_{(2)})$, using $\boldsymbol{\beta}_{(2)}$.
6. Repeat process S times.

After a certain number of iterations, draws will not result from the conditional posterior, but from the unconditional posterior. A large number of draws recovers the unconditional posterior numerically – a key property for modern Bayesian methods, afforded by large developments in computational power in the 21st century. The initial iterations for which the algorithm has not yet reached the unconditional distribution is called the **burn-in sample**, which are usually discarded when constructing the final posterior distribution.

3.3.2 The Metropolis-Hastings algorithm

The MH algorithm is similar to the Gibbs sampler, but there is one main difference: at each iteration, the new draw obtained for the parameter will be accepted only with a certain probability. This probability depends on the conditional density, and if the draw is not accepted, the value from the previous iteration is retained. Also, because we cannot draw directly from the conditional posterior, each new value is drawn from the previous value according to some transition function called the transition kernel. The MH algorithm is roughly as follows:

1. Fix starting values $\boldsymbol{\beta}_{(0)}$ and $\sigma^2_{(0)}$ for $\boldsymbol{\beta}$ and σ^2.
2. Draw a candidate value for $\boldsymbol{\beta}_{(1)}$ from its transition kernel (a function of $\boldsymbol{\beta}_{(0)}$).

3. Compute its acceptance probability from the conditional density, $f(\boldsymbol{\beta}|\mathbf{y}, \sigma^2)$.
4. If the draw is accepted, update the value. Otherwise, keep the former value and set $\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)}$.
5. Draw candidate $\sigma^2_{(1)}$ from its transition kernel (a function of $\sigma^2_{(0)}$).
6. Compute its acceptance probability from the conditional density, $f(\sigma^2|\mathbf{y}, \boldsymbol{\beta}_{(1)})$, using $\boldsymbol{\beta}_{(1)}$.
7. If draw accepted, update the value. Otherwise, set $\sigma^2_{(1)} = \sigma^2_{(0)}$.
8. Repeat process S times.

Similar to the Gibbs sampler, after a certain number of iterations, the algorithm will draw from the unconditional distribution. The MH algorithm is more general than the Gibbs sampler, but heavier in terms of computational requirements. For most econometric applications (such as VARs), the Gibbs sampler is sufficient.

3.3.3 Example: The Gibbs sampler for linear regression models

We obtained the joint posterior distribution, $f(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$, as shown in Equation (23)

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto |\sigma^2\mathbf{I}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \\ \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} (\sigma_0^2)^{-\alpha-1} \exp\left\{-\frac{\beta}{\sigma_0^2}\right\}, \end{aligned}$$

but how do we obtain the conditional distributions $f(\boldsymbol{\beta}|\mathbf{y}, \sigma^2)$ and $f(\sigma^2|\mathbf{y}, \boldsymbol{\beta})$?

Well, hypothetically, if we conditioned $f(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ on σ^2, then it would imply that σ^2 is known and can be treated as a constant. Then, the only remaining argument of the density is $\boldsymbol{\beta}$. Thus, to obtain $f(\boldsymbol{\beta}|\mathbf{y}, \sigma^2)$, we can start from $f(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ and treat any term not involving $\boldsymbol{\beta}$ as part of the proportionality constant (including σ^2). This means that (23) can be written as

$$f(\boldsymbol{\beta}|\mathbf{y}, \sigma^2) \propto \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\},$$

which is exactly the same posterior as (11), where $\boldsymbol{\beta}$ was the only unknown. Like before, $f(\boldsymbol{\beta}|\mathbf{y}, \sigma^2)$ is multivariate normal with mean

$$\bar{\boldsymbol{\beta}} = \bar{\boldsymbol{\Omega}}\left(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y}\right)$$

and variance

$$\bar{\boldsymbol{\Omega}} = \left(\boldsymbol{\Omega}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X}\right)^{-1}.$$

Meanwhile, the conditional posterior of σ^2 (ignoring terms not including σ^2) is

$$f(\sigma^2|\mathbf{y},\boldsymbol{\beta}) \propto |\sigma^2\mathbf{I}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}(\sigma_0^2)^{-\alpha-1}\exp\left\{-\frac{\beta}{\sigma_0^2}\right\},$$
$$\propto (\sigma_0^2)^{-\bar{\alpha}-1}\exp\left\{-\frac{\bar{\beta}}{\sigma_0^2}\right\},$$

where $\bar{\alpha} = \frac{n}{2}+\alpha$ and $\bar{\beta} = \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\beta$. The proportional expression for $f(\sigma^2|\mathbf{y},\boldsymbol{\beta})$ is the kernel of an Inverse-Gamma distribution with shape $\bar{\alpha}$ and scale $\bar{\beta}$.

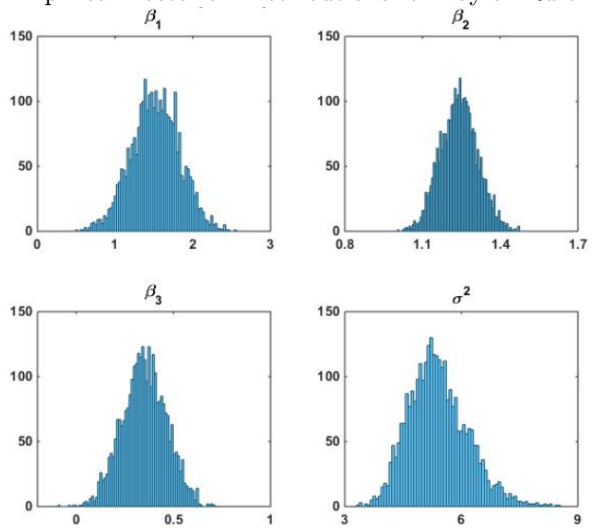
Most software packages will include methods for Bayesian estimation, so while the theory may look slightly daunting, the implementation is not all that hard. The only challenge in Bayesian estimation is finding good data (as is usually the case for most econometric work) and specifying sensible priors.

Consider our estimation of the Taylor Rule, this time with unknown σ^2, parameterised with the Inverse-Gamma distribution with hyperparameters $\alpha = 0.001$ and $\beta = 0.001$:

$$r_t = \underset{(0.89,2.12)}{1.52} + \underset{(1.10,1.38)}{1.24}\bar{\pi}_t + \underset{(0.14,0.59)}{0.35}x_t + \hat{\epsilon}_t,$$

and where the posterior distributions for the parameters are shown in Figure 7. These results should be fairly intuitive: the results are similar to the Bayesian estimate of the simplified model, (17), given that our hyperparameters produce a diffuse prior for σ^2. Thus, the data generates an estimate for σ^2, and the OLS estimate, $\hat{\sigma}^2$, is obtained.

Figure 7: Empirical Posterior Distributions for Taylor Rule Estimation



What if we instead used an Inverse-Gamma distribution with hyperparameters $\alpha = 1000$ and $\beta = 1$

(variance for $\sigma^2 \to 0$)?

$$r_t = \underset{(1.78,2.13)}{1.95} + \underset{(1.11,1.18)}{1.15}\,\bar{\pi}_t + \underset{(0.19,0.31)}{0.25}\,x_t + \hat{\epsilon}_t,$$

which looks very similar to the ML/OLS estimates for $\boldsymbol{\beta}$, (16). This may be surprising at first, but let's consider why this is the case: A loose prior for σ^2 causes a tight prior for $\boldsymbol{\beta}$, and so the posterior for $\boldsymbol{\beta}$ tends towards its prior. Conversely, a tight prior for σ^2 results in a diffuse prior for $\boldsymbol{\beta}$, where the posterior for $\boldsymbol{\beta}$ tends towards the ML/OLS estimate. Remember, σ^2 is the variance for the data. A tight prior for σ^2 means we assume a very small variance for the data. That is, we are very confident in the information provided by the data.

In other words, we tell the model "put all the weight on the likelihood function, and produce ML/OLS estimates for $\boldsymbol{\beta}$". Conversely, a loose prior for σ^2 means we assume a very large variance in the data, implying that we tell the model "put no weight on the data, and instead put all the weight to the prior, so that the posterior for $\boldsymbol{\beta}$ will match the prior".

To summarise what we've covered in this section: for most models, it is not possible to derive an analytical form for the posterior distribution. However, if the conditional posterior is a known distribution, one can use Gibbs sampling to numerically approximate it. If the conditional posterior can be calculated but is not a known distribution, one has to use the MH algorithm instead. The principle behind these MCMC algorithms is that after drawing successively from the conditional posterior, one will eventually draw from the unconditional posterior. Finally, hyperparameters have a very strong impact on the results. If your estimation is failing, or giving you strange results, revise your selections for the hyperparameters.

4 Bayesian Estimation of VAR Models

In the previous section we covered linear regression models using both standard ML and Bayesian techniques. We saw that the choice of prior was important when calculating the posterior. Generally speaking, when we had a rich dataset the posterior was dominated by the dominated by the data, and when we had a noise dataset and a strong prior the posterior was dominated by the prior. We also saw that in some cases it was possible to sample directly from a known posterior distribution. But in cases where no closed form expression of the posterior exists, then we had to resort of simulation (MCMC) methods.

In this section we move onto vector autoregression (VAR) and Bayesian VAR (BVAR) models. For the sake brevity, we won't cover autoregressive (AR) models. Those interested in a deep exposition of AR models can read good texts such as Enders (2010).

4.1 Mathematical prerequisites: VARs

A VAR model is essentially a multivariate version of an AR model, and it can be expressed as

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} a_{10} \\ a_{20} \end{bmatrix} + \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a_{11}^{(p)} & a_{12}^{(p)} \\ a_{21}^{(p)} & a_{22}^{(p)} \end{bmatrix} \begin{bmatrix} x_{t-p} \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} e_{x,t} \\ e_{y,t} \end{bmatrix},$$

or in compact notation as

$$\mathbf{Y}_t = \mathbf{A}_0 + \mathbf{A}_1 \mathbf{Y}_{t-1} + \cdots + \mathbf{A}_p \mathbf{Y}_{t-p} + \mathbf{e}_t, \tag{24}$$

where

$$\mathbf{e}_t \sim \text{IID}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}.$$

Equation (24) captures the co-movement of the variables in $\mathbf{Y}_t$, and here is known as a vector autoregression of order p, or simply VAR(p).

We can introduce more lagged dependent variables, moving average components (giving us a VARMA process), and other explanatory variables too.

4.1.1 Stability and stationarity

Let us focus on the VAR(1) model first:

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{A}_0 + \mathbf{A}_1 \mathbf{Y}_{t-1} + \mathbf{e}_t \\ &= \mathbf{A}_0 + \mathbf{A}_1 \left(\mathbf{A}_0 + \mathbf{A}_1 \mathbf{Y}_{t-1} + \mathbf{e}_{t-1} \right) + \mathbf{e}_t \\ &= (\mathbf{I} + \mathbf{A}_1)\mathbf{A}_0 + \mathbf{e}_t + \mathbf{A}_1 \mathbf{e}_{t-1} + \mathbf{A}_1^2 \mathbf{Y}_{t-2}, \end{aligned} \tag{25}$$

and using backwards recursion, we can derive for the n-th step:

$$\mathbf{Y}_t = (\mathbf{I} + \mathbf{A}_1 + \cdots + \mathbf{A}_1^n)\mathbf{A}_0 + \sum_{i=0}^{n} \mathbf{A}_1^i \mathbf{e}_{t-1} + \mathbf{A}_1^{n+1}\mathbf{Y}_{t-(n+1)}.$$

Similar to the AR(1) process, if $\mathbf{A}_1^{n+1}$ approaches the null matrix, then we can write our VAR(1) process as the following infinite vector moving average (VMA) process:

$$\mathbf{Y}_t = \sum_{i=0}^{\infty}(\mathbf{A}_1^i)\mathbf{A}_0 + \sum_{i=0}^{\infty} \mathbf{A}_1^i \mathbf{e}_{t-i}. \tag{26}$$

Recall our stability conditions when we looked at DSGE models (and VAR models) in first-year macroeconomics, and consider the singular value decomposition of the $\mathbf{A}_1$ matrix:[2]

$$\begin{aligned}
\mathbf{A}_1 &= \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top, \\
\mathbf{A}_1^2 &= \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top = \mathbf{U}\mathbf{\Lambda}^2\mathbf{V}^\top, \\
&\vdots \\
\mathbf{A}_1^n &= \mathbf{U}\mathbf{\Lambda}^n\mathbf{V}^\top,
\end{aligned}$$

where $\mathbf{\Lambda}$ is the diagonal matrix with singular values, i.e.,

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

and $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices whose columns are the eigenvectors of $\mathbf{A}_1$. Therefore, if $|\lambda_1|, |\lambda_2| < 1$, then $\mathbf{A}_1^n \to 0$ as $n \to +\infty$, and the system is stable.

From Equation (26), we derive that

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}_t] &= \sum_{i=0}^{\infty}(\mathbf{A}_1^i)\mathbf{A}_0 = (\mathbf{I} - \mathbf{A}_1)^{-1}\mathbf{A}_0, \\
\mathrm{Var}(\mathbf{Y}_t) &= \mathrm{Var}\left(\sum_{i=0}^{\infty} \mathbf{A}_1^i \mathbf{e}_{t-i}\right) = \mathbf{\Sigma} + \mathbf{A}_1\mathbf{\Sigma}\mathbf{A}_1^\top + \mathbf{A}_1^2\mathbf{\Sigma}(\mathbf{A}_1^2)^\top + \cdots.
\end{aligned}$$

One can show that

$$\mathrm{vec}\left(\mathrm{Var}(\mathbf{Y}_t)\right) = (\mathbf{I} - \mathbf{A}_1 \otimes \mathbf{A}_1)^{-1}\mathrm{vec}\left(\mathrm{Var}(\mathbf{e}_t)\right),$$

where $\otimes$ is the Kronecker product. We have not derived the autocovariance function, but we can

[2] Excuse the abuse of notation with respect to the powers and transposes. The point I want to make here is the eigenvalue decomposition.

observe that if the eigenvalues of $\mathbf{\Lambda}$ are within unit circle, then the variables in $\mathbf{Y}_t$ are jointly covariance stationary.

In general, any VAR(p) process has a VAR(1) representation in its mean adjusted form:

$$\mathbf{Y}_t - \boldsymbol{\mu} = \mathbf{A}_1(\mathbf{Y}_{t-1} - \boldsymbol{\mu}) + \cdots + \mathbf{A}_p(\mathbf{Y}_{t-p} - \boldsymbol{\mu}) + \mathbf{e}_t, \tag{27}$$

where $\mu = (\mathbf{I} - \mathbf{A}_1 - \cdots - \mathbf{A}_p)^{-1}\mathbf{A}_0$. We can rewrite (27) as

$$\mathbf{Z}_t = \mathbf{F}\mathbf{Z}_{t-1} + \mathbf{E}_t,$$

where

$$\mathbf{Z}_t = \begin{bmatrix} \mathbf{Y}_t - \boldsymbol{\mu} \\ \vdots \\ \mathbf{Y}_{t-p+1} - \boldsymbol{\mu} \end{bmatrix}, \ \mathbf{F} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_p \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix}, \ \mathbf{E}_t = \begin{bmatrix} \mathbf{e}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$$

The matrix $\mathbf{F}$ is known as the companion matrix. The stability restrictions are given by the eigenvalues of $\mathbf{F}$. If its eigenvalues are less than unity in absolute value, then the system is stable and therefore is covariance stationary.

4.1.2 Estimation, inference, and lag selection

The VAR(p) model in Equation (24) can be rewritten as

$$\mathbf{Y}_t^\top = \mathbf{A}_0^\top + \mathbf{Y}_{t-1}^\top\mathbf{A}_1^\top + \cdots + \mathbf{Y}_{t-p}\mathbf{A}_p^\top + \mathbf{e}_t^\top, \tag{28}$$

or, in matrix notation,

$$\begin{aligned} \mathbf{Y}_t^\top &= \begin{bmatrix} 1 & \mathbf{Y}_{t-1}^\top & \cdots & \mathbf{Y}_{t-p}^\top \end{bmatrix} \begin{bmatrix} \mathbf{A}_0^\top \\ \mathbf{A}_1^\top \\ \vdots \\ \mathbf{A}_p^\top \end{bmatrix} + \mathbf{e}_t^\top \\ &= \mathbf{X}_t\mathbf{\Pi} + \mathbf{e}_t^\top, \end{aligned}$$

where $\mathbf{Y}_t^\top$ is a $1 \times g$ row vector.[3] So the VAR(p) model has the same representation as a simple linear regression model:

$$\underset{T\times g}{\mathbf{Y}} = \underset{T\times(p+1)g}{\mathbf{X}} \underset{(p+1)g\times g}{\mathbf{\Pi}} + \underset{T\times g}{\mathbf{e}}, \tag{29}$$

[3] A lot of people get confused here due to different textbooks/lecturers using different matrix dimensions and notation. Here the matrix $\mathbf{Y}$ has n observations, with g explanatory variables (not including the intercept term), and p lags. More

and the parameters of the VAR model can be estimated by OLS:

$$\hat{\mathbf{\Pi}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}. \tag{30}$$

As before we have

$$\begin{aligned} \hat{\mathbf{\Pi}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\mathbf{\Pi} + \mathbf{e}) \\ &= \mathbf{\Pi} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{e}. \end{aligned}$$

Using the vec operator, we can stack the columns of the $(p+1)g \times g$ matrix $\hat{\mathbf{\Pi}}$ into a long $(p+1)g^2 \times 1$ vector:

$$\text{vec}\left(\hat{\mathbf{\Pi}}\right) = \text{vec}(\mathbf{\Pi}) + \text{vec}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{e}\right). \tag{31}$$

One can show that the variance of $\text{vec}\left(\hat{\mathbf{\Pi}}\right)$ is

$$\text{Var}\left(\text{vec}\left(\hat{\mathbf{\Pi}}\right)\right) = \mathbf{\Sigma} \otimes (\mathbf{X}^\top \mathbf{X})^{-1}$$

An estimator for $\mathbf{\Sigma}$ is given by

$$\begin{aligned} \hat{\mathbf{\Sigma}} &= \frac{1}{T} \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}}\right)^\top \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}}\right) \\ &= \frac{1}{T} \hat{\mathbf{e}}^\top \hat{\mathbf{e}}. \end{aligned}$$

$\hat{\mathbf{\Pi}}$ and $\hat{\mathbf{\Sigma}}$ are the MLEs of $\mathbf{\Pi}$ and $\mathbf{\Sigma}$, respectively. These estimators are obtained from maximising the log likelihood function,

$$-\frac{Tg}{2} \log 2\pi - \frac{T}{2} \log |\mathbf{\Sigma}| - \frac{1}{2} \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{\Pi})^\top (\mathbf{\Sigma}^{-1} \otimes \mathbf{I}) \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{\Pi}).$$

We can perform a t-test for any parameter of the VAR model by selecting the appropriate diagonal element of $\hat{\mathbf{\Sigma}} \otimes (\mathbf{X}^\top \mathbf{X})^{-1}$. Additionally, we can test several parameters simultaneously – e.g., let $\hat{\boldsymbol{\pi}}_{p+1}$

generally, we could also include exogenous variables, $\mathbf{V}_t$, to have

$$\begin{aligned} \mathbf{Y}_t^\top &= \mathbf{A}_{01}^\top + \mathbf{V}_t^\top \mathbf{A}_{02}^\top + \mathbf{Y}_{t-1}^\top \mathbf{A}_1^\top + \cdots + \mathbf{Y}_{t-p} \mathbf{A}_p^\top + \mathbf{e}_t^\top, \\ \Leftrightarrow \mathbf{Y}_t^\top &= \begin{bmatrix} 1 & \mathbf{V}_t & \mathbf{Y}_{t-1}^\top & \cdots & \mathbf{Y}_{t-p}^\top \end{bmatrix} \begin{bmatrix} \mathbf{A}_{01}^\top \\ \mathbf{A}_{02}^\top \\ \mathbf{A}_1^\top \\ \vdots \\ \mathbf{A}_p^\top \end{bmatrix} + \mathbf{e}_t^\top, \end{aligned}$$

or, alternatively,

$$\underset{T \times g}{\mathbf{Y}} = \underset{T \times (p+2)g}{\mathbf{X}} \underset{(p+2)g \times g}{\mathbf{\Pi}} + \underset{T \times g}{\mathbf{e}}.$$

and $\widehat{\text{Var}}(\hat{\boldsymbol{\pi}}_{p+1})$ be the estimators of $\text{vec}(\mathbf{A}_{p+1})$ and $\text{Var}\left(\text{vec}\left(\hat{\mathbf{A}}_{p+1}\right)\right)$, respectively. Then

$$(\hat{\boldsymbol{\pi}}_{p+1} - \boldsymbol{\pi}_{p+1})^\top \widehat{\text{Var}}(\hat{\boldsymbol{\pi}}_{p+1})^{-1} (\hat{\boldsymbol{\pi}}_{p+1} - \boldsymbol{\pi}_{p+1}) \sim \chi^2(g^2).$$

If we assume that $H_0 : \mathbf{A}_{p+1} = \mathbf{O}$, then the above statistic becomes

$$\hat{\boldsymbol{\pi}}_{p+1}^\top \widehat{\text{Var}}(\hat{\boldsymbol{\pi}}_{p+1}) \hat{\boldsymbol{\pi}}_{p+1} \sim \chi^2(g^2).$$

4.1.3 Lag selection

The number of parameters in Equation (28) is $(p+1)g^2 + \frac{g(g+1)}{2}$, and an extra lag would add another g^2 coefficients. So it is very important to determine the number of lags in the model. One option is to use the following likelihood ratio test:

$$T\left(\log\left|\hat{\boldsymbol{\Sigma}}(p)\right| - \log\left|\hat{\boldsymbol{\Sigma}}(p+1)\right|\right) \sim \chi^2(g^2), \tag{32}$$

where $\hat{\boldsymbol{\Sigma}}(p)$ and $\hat{\boldsymbol{\Sigma}}(p+1)$ are the MLEs of $\boldsymbol{\Sigma}$ for VAR(p) and VAR($p+1$) models. Under the null, $H_0 : \mathbf{A}_{p+1} = \mathbf{O}$, we are testing if the last coefficients are statistically significant.

We can also use the Akaike Information Criterion (AIC) and/or Schwarz-Bayesian Information Criterion (BIC) for choosing the optimal number of lags:

$$AIC = T\log\left|\hat{\boldsymbol{\Sigma}}\right| + 2m,$$
$$BIC = T\log\left|\hat{\boldsymbol{\Sigma}}\right| + m\log T,$$

where m is the total number of parameters estimated in the VAR, and $\left|\hat{\boldsymbol{\Sigma}}\right|$ is the determinant of $\hat{\boldsymbol{\Sigma}}$.

4.1.4 Granger causality

Let us divide the elements of the VAR(p) model into two groups as follows

$$\mathbf{Y}_t^\top = \begin{bmatrix}\mathbf{Y}_{1,t}^\top & \mathbf{Y}_{2,t}^\top\end{bmatrix},$$

where $\mathbf{Y}_{1,t}^\top$ and $\mathbf{Y}_{2,t}^\top$ are $1 \times g_1$ and $1 \times g_2$ row vectors, respectively, with $g_1 + g_2 = g$:

$$\begin{bmatrix}\mathbf{Y}_{1,t}\\ \mathbf{Y}_{2,t}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{10}\\ \mathbf{A}_{20}\end{bmatrix} + \begin{bmatrix}\mathbf{A}_{11}^{(1)} & \mathbf{A}_{12}^{(1)}\\ \mathbf{A}_{21}^{(1)} & \mathbf{A}_{22}^{(1)}\end{bmatrix}\begin{bmatrix}\mathbf{Y}_{1,t-1}\\ \mathbf{Y}_{2,t-1}\end{bmatrix} + \cdots + \begin{bmatrix}\mathbf{A}_{11}^{(p)} & \mathbf{A}_{12}^{(p)}\\ \mathbf{A}_{21}^{(p)} & \mathbf{A}_{22}^{(p)}\end{bmatrix}\begin{bmatrix}\mathbf{Y}_{1,t-p}\\ \mathbf{Y}_{2,t-p}\end{bmatrix} + \begin{bmatrix}\mathbf{e}_{1,t}\\ \mathbf{e}_{2,t}\end{bmatrix}.$$

If $\mathbf{Y}_{2,t}$ does not Granger cause $\mathbf{Y}_{1,t}$, then $\mathbf{A}_{12}^{(1)} = \cdots = \mathbf{A}_{12}^{(p)} = \mathbf{O}$.[4] Granger causality refers only to the effects of the past values of $\mathbf{Y}_{2,t}$ on the current values of $\mathbf{Y}_{1,t}$. Granger causality does not exclude the contemporaneous effects of $\mathbf{Y}_{1,t}$ on $\mathbf{Y}_{2,t}$ which is given by the correlation between $\mathbf{e}_{1,t}$ and $\mathbf{e}_{2,t}$.

We can test Granger causality by running an F-test. The restricted and unrestricted models are:

$$\mathbf{Y}_{1,t} = \mathbf{A}_{10} + \sum_{i=1}^{p} \mathbf{A}_{11}^{(i)} \mathbf{Y}_{1,t-i} + \mathbf{e}_{1,t}, \tag{33}$$

$$\mathbf{Y}_{2,t} = \mathbf{A}_{10} + \sum_{i=1}^{p} \left(\mathbf{A}_{11}^{(i)} \mathbf{Y}_{1,t-i} + \mathbf{A}_{12}^{(i)} \mathbf{Y}_{2,t-i} \right) + \mathbf{e}_{1,t}. \tag{34}$$

Alternatively, we can compute the likelihood ratio test similar to the test (32):

$$T \left(\log \left| \tilde{\mathbf{\Sigma}}_{11} \right| - \log \left| \hat{\mathbf{\Sigma}}_{11} \right| \right) \sim \chi^2(pg_1g_2),$$

where $\tilde{\mathbf{\Sigma}}_{11}$ and $\hat{\mathbf{\Sigma}}_{11}$ denote the estimates of $\mathbf{\Sigma}_{11}$ based on the residuals of equations (33) and (34), respectively.

4.1.5 Impulse response functions

Consider the following VAR(1) process:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} a_{10} \\ a_{20} \end{bmatrix} + \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} e_{x,t} \\ e_{y,t} \end{bmatrix},$$

or, in compact notation,

$$\mathbf{Y}_t = \mathbf{A}_0 + \mathbf{A}_1 \mathbf{Y}_{t-1} + \mathbf{e}_t,$$

where

$$\mathbf{e}_t \sim \text{IID}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix},$$

We know that we can write any VAR(p) process as an infinite-order VMA process – i.e., here we can write the VAR(1) model as

$$\mathbf{Y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \mathbf{A}_1^i \mathbf{e}_{t-i},$$

where $\boldsymbol{\mu} = \sum_{i=0}^{\infty} (\mathbf{A}_1^i) \mathbf{A}_0$.

[4] Similarly, if $\mathbf{Y}_{1,t}$ does not Granger cause $\mathbf{Y}_{2,t}$, then $\mathbf{A}_{21}^{(1)} = \cdots = \mathbf{A}_{21}^{(p)} = \mathbf{O}$.

Let us consider the following Cholesky decomposition of the variance covariance matrix $\mathbf{\Sigma}$:

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{P}\mathbf{P}^\top \\ \implies \mathbf{P}^{-1}\mathbf{\Sigma}(\mathbf{P}^\top)^{-1} &= \mathbf{I}, \end{aligned}$$

where $\mathbf{P}$ is a lower triangular matrix. So we have

$$\begin{aligned} \mathbf{Y}_t &= \mu + \sum_{i=0}^{\infty} \mathbf{A}_1^i \mathbf{P}\mathbf{P}^{-1}\mathbf{e}_{t-i} \\ &= \mu + \sum_{i=0}^{\infty} \mathbf{\Phi}(i)\boldsymbol{\varepsilon}_{t-i}, \end{aligned} \tag{35}$$

where $\mathbf{\Phi}(i) = \mathbf{A}_1^i\mathbf{P}$ and $\boldsymbol{\varepsilon}_{t-i} = \mathbf{P}^{-1}\mathbf{e}_{t-i}$. Note that the variance of $\boldsymbol{\varepsilon}_t$ is

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top\right] &= \mathbb{E}\left[\mathbf{P}^{-1}\mathbf{e}_{t-i}\mathbf{e}_{t-i}^\top(\mathbf{P}^\top)^{-1}\right] \\ &= \mathbf{P}^{-1}\mathbb{E}\left[\mathbf{e}_{t-i}\mathbf{e}_{t-i}^\top\right](\mathbf{P}^\top)^{-1} \\ &= \mathbf{P}^{-1}\mathbf{\Sigma}(\mathbf{P}^\top)^{-1} \\ &= \mathbf{I}. \end{aligned}$$

Examining (35) further, we can see

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \phi_{11}^{(i)} & \phi_{12}^{(1)} \\ \phi_{21}^{(1)} & \phi_{22}^{(1)} \end{bmatrix} \begin{bmatrix} \varepsilon_{x,t-1} \\ \varepsilon_{y,t-1} \end{bmatrix},$$

where $\phi_{12}^{(0)}$ captures the effect of a one unit change of $\varepsilon_{y,t}$ on x_t, $\phi_{12}^{(1)}$ represents the impact of one unit of $\varepsilon_{y,t-1}$ ($\varepsilon_{y,t}$) on x_t (x_{t+1}), and so on. All the coefficients of $\mathbf{\Phi}(i)$ are known as the impulse response functions (IRFs). Usually we plot the coefficients $\phi_{kj}^{(i)}$ against i to visualise the impact of a shock on the paths of the variables in $\mathbf{Y}_t$.

The cumulated sum of the effects of the shocks of $\varepsilon_{y,t}$ on x_t is

$$C(n) = \sum_{i=0}^{n} \phi_{12}^{(n)}.$$

Also, consider our simple VAR(1) model again:

$$\begin{aligned}
\mathbf{Y}_t &= \mathbf{A}_0 + \mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{e}_t \\
\mathbf{P}^{-1}\mathbf{Y}_t &= \mathbf{P}^{-1}\mathbf{A}_0 + \mathbf{P}^{-1}\mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{P}^{-1}\mathbf{e}_t \\
\mathbf{P}^{-1}\mathbf{Y}_t + \mathbf{Y}_t &= \mathbf{Y}_t + \mathbf{P}^{-1}\mathbf{A}_0 + \mathbf{P}^{-1}\mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{P}^{-1}\mathbf{e}_t \\
\mathbf{Y}_t &= \underbrace{(\mathbf{I} - \mathbf{P}^{-1})}_{\mathbf{D}}\mathbf{Y}_t + \mathbf{P}^{-1}\mathbf{A}_0 + \mathbf{P}^{-1}\mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{P}^{-1}\mathbf{e}_t.
\end{aligned}$$

Clearly, $\mathbf{D}$ is a lower triangular matrix. The Cholesky decomposition imposes a recursive casual structure from the top variables to the bottom variables but not on the other way around. Therefore, the IRF is sensitive to the order of variables in $\mathbf{Y}_t$.

4.1.6 Forecast error variance decomposition

Consider again the simple VAR(1) model:

$$\begin{aligned}
\mathbf{Y}_t &= \mathbf{A}_0 + \mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{e}_t \\
\implies \mathbf{Y}_{t+1} &= \mathbf{A}_0 + \mathbf{A}_1\mathbf{Y}_t + \mathbf{e}_{t+1},
\end{aligned}$$

and suppose we would like to predict the value of the variables at $t+1$, but we only have information up to t. At $t+1$ we have

$$\mathbb{E}_t[\mathbf{Y}_{t+1}] = \mathbf{A}_0 + \mathbf{A}_1\mathbf{Y}_t,$$

where $\mathbb{E}_t$ is the conditional expectation operator. The one-step-ahead forecast error is

$$\mathbf{Y}_{t+1} - \mathbb{E}_t[\mathbf{Y}_{t+1}] = \mathbf{e}_{t+1}.$$

Using backwards substitution, we can write $\mathbf{Y}_{t+2}$ as

$$\begin{aligned}
\mathbf{Y}_{t+2} &= \mathbf{A}_0 + \mathbf{A}_1(\underbrace{\mathbf{A}_0 + \mathbf{A}_1\mathbf{Y}_t + \mathbf{e}_{t+1}}_{\mathbf{Y}_{t+1}}) + \mathbf{e}_{t+2} \\
&= \underbrace{(\mathbf{I} + \mathbf{A}_1)\mathbf{A}_0 + \mathbf{A}_1^2\mathbf{Y}_t}_{\mathbb{E}_t[\mathbf{Y}_{t+2}]} + \mathbf{e}_{t+2} + \mathbf{A}_1\mathbf{e}_{t+1},
\end{aligned}$$

with the forecast error

$$\mathbf{Y}_{t+2} - \mathbb{E}_t[\mathbf{Y}_{t+2}] = \mathbf{e}_{t+2} + \mathbf{A}_1\mathbf{e}_{t+1}.$$

Extending the procedure for the n-th step, we have

$$\mathbf{Y}_{t+n} = \underbrace{(\mathbf{I} + \mathbf{A}_1 + \cdots + \mathbf{A}_1^{n-1})\mathbf{A}_0 + \mathbf{A}_1^n \mathbf{Y}_t}_{\mathbb{E}_t[\mathbf{Y}_{t+n}]} + \sum_{i=0}^{n-1} \mathbf{A}_1^i \mathbf{e}_{t+n-i},$$

with the associated forecast error of

$$\mathbf{Y}_{t+n} - \mathbb{E}_t[\mathbf{Y}_{t+n}] = \sum_{i=0}^{n-1} \mathbf{A}_1^i \mathbf{e}_{t+n-i}.$$

Using the Cholesky decomposition from (35), we can rewrite the above equation as

$$\mathbf{Y}_{t+n} - \mathbb{E}_t[\mathbf{Y}_{t+n}] = \sum_{i=0}^{n-1} \mathbf{\Phi}(i)\boldsymbol{\varepsilon}_{t+n-i}$$

Thus, the variance of the n-step ahead forecast error is

$$\begin{aligned}
\text{Var}\left(\mathbf{Y}_{t+n} - \mathbb{E}_t[\mathbf{Y}_{t+n}]\right) &= \text{Var}\left(\sum_{i=0}^{n-1} \mathbf{\Phi}(i)\boldsymbol{\varepsilon}_{t+n-i}\right) \\
&= \sum_{i=0}^{n-1} \text{Var}\left(\mathbf{\Phi}(i)\boldsymbol{\varepsilon}_{t+n-i}\right) \\
&= \sum_{i=0}^{n-1} \mathbf{\Phi}(i)\text{Var}\left(\boldsymbol{\varepsilon}_{t+n-i}\right)\mathbf{\Phi}(i)^\top.
\end{aligned}$$

Looking into further detail of this matrix, we have

$$\sum_{i=0}^{n-1} \begin{bmatrix} \phi_{11}^{(i)} & \phi_{12}^{(i)} \\ \phi_{21}^{(i)} & \phi_{22}^{(i)} \end{bmatrix} \begin{bmatrix} \sigma_{\varepsilon_x}^2 & 0 \\ 0 & \sigma_{\varepsilon_y}^2 \end{bmatrix} \begin{bmatrix} \phi_{11}^{(i)} & \phi_{21}^{(i)} \\ \phi_{12}^{(i)} & \phi_{22}^{(i)} \end{bmatrix} = \sum_{i=0}^{n-1} \begin{bmatrix} (\phi_{11}^{(i)})^2\sigma_{\varepsilon_x}^2 (\phi_{12}^{(i)})^2\sigma_{\varepsilon_y}^2 & \phi_{11}^{(i)}\phi_{21}^{(i)}\sigma_{\varepsilon_x}^2 + \phi_{12}^{(i)}\phi_{22}^{(i)}\sigma_{\varepsilon_y}^2 \\ \phi_{11}^{(i)}\phi_{21}^{(i)}\sigma_{\varepsilon_x}^2 + \phi_{12}^{(i)}\phi_{22}^{(i)}\sigma_{\varepsilon_y}^2 & (\phi_{21}^{(i)})^2\sigma_{\varepsilon_x}^2 + (\phi_{22}^{(i)})^2\sigma_{\varepsilon_y}^2 \end{bmatrix}.$$

Let us focus on the variance of the forecast error, $\sigma_x^2(n) = \text{Var}\left(x_{t+n} - \mathbb{E}_t[x_{t+n}]\right)$,

$$\sigma_x^2(n) = \sum_{i=0}^{n-1} (\phi_{11}^{(i)})^2\sigma_{\varepsilon_x}^2 + (\phi_{12}^{(i)})^2\sigma_{\varepsilon_y}^2.$$

The proportion of $\sigma_x^2(n)$ due to shocks on $\varepsilon_{x,t}$ and $\varepsilon_{y,t}$ are, respectively,

$$\frac{\sum_{i=0}^{n-1}(\phi_{11}^{(i)})^2\sigma_{\varepsilon_x}^2}{\sigma_x^2(n)}, \text{ and}$$
$$\frac{\sum_{i=0}^{n-1}(\phi_{12}^{(i)})^2\sigma_{\varepsilon_y}^2}{\sigma_x^2(n)}.$$

If $(\phi_{12}^{(i)})^2 = 0$ for all i, we say that the variable x_t is exogenous (it does not depend on either the $\varepsilon_{y,t}$ shocks nor on the y_t sequence).

4.1.7 Structural VAR

Consider the following VAR(1) process

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \end{bmatrix} + \begin{bmatrix} 0 & b_{12} \\ b_{21} & 0 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix}, \tag{36}$$

where

$$\begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix} \sim \text{IID}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\varepsilon_x}^2 & 0 \\ 0 & \sigma_{\varepsilon_y}^2 \end{bmatrix} \right).$$

We can rewrite (36) as

$$\begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix},$$

or simply

$$\mathbf{B}\mathbf{Y}_t = \mathbf{\Gamma}_0 + \mathbf{\Gamma}_1\mathbf{Y}_{t-1} + \boldsymbol{\varepsilon}_t. \tag{37}$$

Premultiplying both sides by $\mathbf{B}^{-1}$, we obtain

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{B}^{-1}\mathbf{\Gamma}_0 + \mathbf{B}^{-1}\mathbf{\Gamma}_1\mathbf{Y}_{t-1} + \mathbf{B}^{-1}\boldsymbol{\varepsilon}_t \\ \Leftrightarrow \mathbf{Y}_t &= \mathbf{A}_0 + \mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{e}_t, \end{aligned} \tag{38}$$

where obviously $\mathbf{A}_0 = \mathbf{B}^{-1}\mathbf{\Gamma}_0$, $\mathbf{A}_1 = \mathbf{B}^{-1}\mathbf{\Gamma}_1$, and $\mathbf{e}_t = \mathbf{B}^{-1}\boldsymbol{\varepsilon}_t$. The system (37) is known as a structural VAR (SVAR) model, while the system (38) is its reduced form representation, which is exactly what we had in (24) and (25).

We have so far discussed estimation and inference of reduced form VARs. However, we are interested in the structural parameters. It makes more sense to derive the IRFs and the forecast variance decomposition from the structural model instead of the reduced form model. In the example of Equation (36), there are 10 structural parameters, but we can only estimate 9 reduced form parameters

from the reduced form representation in Equation (38) – i.e., without imposing any restriction on the structural system, we will not be able to estimate the structural parameters from the reduced form model.

In general, consider the following SVAR(p) model

$$\mathbf{B}\mathbf{Y}_t = \mathbf{\Gamma}_0 + \mathbf{\Gamma}_1\mathbf{Y}_{t-1} + \cdots + \mathbf{\Gamma}_p\mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{Y}_t$ is a $g\times 1$ vector. So there are (g^2-g) structural parameters in matrix $\mathbf{B}$ (as the diagonal terms are ones), $(g+pg^2)$ elements in the $\mathbf{\Gamma}$ matrices, and g variance terms in $\text{Var}(\boldsymbol{\varepsilon}_t)$.[5] The corresponding reduced form VAR is

$$\mathbf{Y}_t = \mathbf{A}_0 + \mathbf{A}_1\mathbf{Y}_{t-1} + \cdots + \mathbf{A}_p\mathbf{Y}_{t-p} + \mathbf{e}_t,$$

where we have $(g+pg^2)$ elements in the $\mathbf{A}$'s and $\frac{g+g^2}{2}$ in $\mathbf{\Sigma}$. Therefore, we need to impose the following number of restrictions:

$$\underbrace{(g^2-g)+(g+pg^2)+g}_{\#\text{ of structural parameters}} - \underbrace{(g+pg^2)+\frac{g+g^2}{2}}_{\#\text{ of reduced form parameters}} = \frac{g^2-g}{2}.$$

As an example, asserting that the reduced form VAR is the structural model is the same as imposing the $\frac{g^2-g}{2}$ a priori restrictions that $\mathbf{A}=\mathbf{I}$.

SVARs generally identify their shocks as coming from distinct independent sources and thus assume that they are uncorrelated. The error series in reduced form VARs are usually correlated with each other. One way to view these correlations is that the reduced from errors are combinations of a set of statistically independent structural errors. The most popular SVAR method is the recursive identification method. This method (used in the original Sims (1980) paper) uses simple regression techniques to construct a set of uncorrelated structural shocks directly from the reduced form shocks. This method sets $\mathbf{A}=\mathbf{I}$ and creates a structure for the shocks to be uncorrelated.

[5] More generally, we could have an SVAR(p) system such as

$$\mathbf{B}\mathbf{Y}_t = \mathbf{\Gamma}_0 + \mathbf{\Gamma}_1\mathbf{Y}_{t-1} + \cdots + \mathbf{\Gamma}_p\mathbf{Y}_{t-p} + \mathbf{C}\boldsymbol{\varepsilon}_t,$$

where we would have (g^2-g) structural parameters in $\mathbf{B}$, $(g+pg^2)$ in the $\mathbf{\Gamma}$'s, g^2 in $\mathbf{C}$, and $\frac{(g+g^2)}{2}$ parameters in $\text{Var}(\boldsymbol{\varepsilon}_t)$.

4.2 Bayes' Rule for VARs

With all the prerequisites out of the way, we are ready to tackle BVARs. The notation from here on out is going to be messy, so let's stick to a simple example. Consider (25) the simple case where $g = 2$,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{10} \\ a_{20} \end{bmatrix} + \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix},$$

which, following what we did in Section 4.1.2, we could also rewrite as[6]

$$\mathbf{Y}_t^\top = \mathbf{A}_0^\top + \mathbf{Y}_{t-1}^\top \mathbf{A}_1^\top + \mathbf{e}_t^\top,$$

where

$$\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}.$$

We can also write our simple system in matrix notation as

$$\mathbf{Y}_t^\top = \underbrace{\begin{bmatrix} 1 & \mathbf{Y}_{t-1}^\top \end{bmatrix}}_{\mathbf{X}_t} \underbrace{\begin{bmatrix} \mathbf{A}_0^\top \\ \mathbf{A}_1^\top \end{bmatrix}}_{\mathbf{\Pi}} + \mathbf{e}_t^\top, \tag{39}$$

so $\mathbf{Y}_t^\top$ is a $1 \times g$ row vector. Then we can write the VAR(1) model has the same representation as the seemingly unrelated regression (SUR) linear regression form:

$$\underset{T\times 2}{\mathbf{Y}} = \underset{T\times(1+1)2(1+1)2\times 2}{\mathbf{X}} \underset{}{\mathbf{\Pi}} + \underset{T\times 2}{\mathbf{e}}, \tag{40}$$

and via OLS/ML we have:

$$\begin{aligned} \hat{\mathbf{\Pi}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\mathbf{\Pi} + \mathbf{e}) \\ &= \mathbf{\Pi} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{e}. \end{aligned}$$

Now, recall the rules of the vec operator:

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\text{vec}(\mathbf{B}),$$

[6] We can also easily add more lags, e.g.,

$$\mathbf{Y}_t^\top = \mathbf{A}_0^\top + \mathbf{Y}_{t-1}^\top \mathbf{A}_1^\top + \cdots + \mathbf{Y}_{t-p}^\top \mathbf{A}_p^\top + \mathbf{e}_t^\top.$$

and its corollary:

$$\begin{aligned}\text{vec}(\mathbf{AB}) &= (\mathbf{I} \otimes \mathbf{A})\text{vec}(\mathbf{B}) \\ &= (\mathbf{B}^\top \otimes \mathbf{I})\text{vec}(\mathbf{A}).\end{aligned}$$

Then use the vec operator on the $(p+1)g \times g$ matrix $\hat{\mathbf{\Pi}}$ and turn it into a long $(p+1)g^2 \times 1$ vector:

$$\text{vec}\left(\hat{\mathbf{\Pi}}\right) = \text{vec}\left(\mathbf{\Pi}\right) + \text{vec}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e}\right),$$

where we can show

$$\text{Var}\left(\text{vec}\left(\hat{\mathbf{\Pi}}\right)\right) = \mathbf{\Sigma} \otimes (\mathbf{X}^\top\mathbf{X})^{-1},$$

which implies

$$\begin{aligned}\hat{\mathbf{\Sigma}} &= \frac{1}{T-(p+1)g}\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}}\right)^\top\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}}\right) \\ &= \frac{1}{T-(p+1)g}\hat{\mathbf{e}}^\top\hat{\mathbf{e}}.\end{aligned}$$

As shown before, $\hat{\mathbf{\Pi}}$ and $\hat{\mathbf{\Sigma}}$ are the MLEs of $\mathbf{\Pi}$ and $\mathbf{\Sigma}$, respectively, and are obtained by maximising the log likelihood function,

$$\begin{aligned}l(\mathbf{Y}|\mathbf{\Pi},\mathbf{\Sigma}) &= -\frac{Tg}{2}\ln(2\pi) - \frac{T}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\text{vec}\left(\mathbf{Y}-\mathbf{X\Pi}\right)^\top\left(\mathbf{I}\otimes\mathbf{\Sigma}\right)^{-1}\text{vec}\left(\mathbf{Y}-\mathbf{X\Pi}\right) \\ \implies f(\mathbf{Y}|\mathbf{\Pi},\mathbf{\Sigma}) &= (2\pi)^{-\frac{Tg}{2}}|\mathbf{\Sigma}|^{-\frac{T}{2}}\exp\left\{\frac{1}{2}\text{vec}\left(\mathbf{Y}-\mathbf{X\Pi}\right)^\top\left(\mathbf{I}\otimes\mathbf{\Sigma}\right)^{-1}\text{vec}\left(\mathbf{Y}-\mathbf{X\Pi}\right)\right\} \\ \Leftrightarrow f(\mathbf{Y}|\mathbf{\Pi},\mathbf{\Sigma}) &= (2\pi)^{-\frac{Tg}{2}}|\mathbf{\Sigma}|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})^\top(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})\right)\right\} \\ &\quad\times\exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{\Pi}-\hat{\mathbf{\Pi}})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{\Pi}-\hat{\mathbf{\Pi}})\right)\right\}. \qquad (41)\end{aligned}$$

Just for exposition, let's look at the representation of our VAR model in Equation (39) again, with

$p = 1$, $g = 2$, and $t = 1, ..., T$, in vector form. If we vectorise $\mathbf{\Pi}$ we get:

$$\begin{aligned}\text{vec}(\mathbf{\Pi}) &= \text{vec}\left(\begin{bmatrix}\mathbf{A}_0^\top \\ \mathbf{A}_1^\top\end{bmatrix}\right) \\ &= \begin{bmatrix}a_{10} \\ a_{11}^{(1)} \\ a_{12}^{(1)} \\ a_{20} \\ a_{21}^{(1)} \\ a_{22}^{(1)}\end{bmatrix} = \boldsymbol{\beta}.\end{aligned}$$

With $\boldsymbol{\beta}$ in hand, we can then rewrite our VAR(1) system as for period t as:

$$\begin{bmatrix}y_{1,t} \\ y_{2,t}\end{bmatrix} = \begin{bmatrix}1 & y_{1,t-1} & y_{2,t-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & y_{1,t-1} & y_{2,t-2}\end{bmatrix} \begin{bmatrix}a_{10} \\ a_{11}^{(1)} \\ a_{12}^{(1)} \\ a_{20} \\ a_{21}^{(1)} \\ a_{22}^{(1)}\end{bmatrix} + \begin{bmatrix}e_{1,t} \\ e_{2,t}\end{bmatrix}$$

$$\Leftrightarrow \mathbf{Y}_t = (\mathbf{I}_2 \otimes \mathbf{X}_t)\boldsymbol{\beta} + \mathbf{e}_t.$$

Here is where our notation can get a bit awkward. Let's define

$$\begin{aligned}\mathbf{y} = \text{vec}(\mathbf{Y}) &= \begin{bmatrix}\mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_T\end{bmatrix}, \\ \mathbf{x}_t &= \mathbf{I}_g \otimes \mathbf{X}_t, \\ \mathbf{u} = \text{vec}(\mathbf{e}) &= \begin{bmatrix}\mathbf{e}_1 \\ \vdots \\ \mathbf{e}_T\end{bmatrix},\end{aligned}$$

and so we have

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u}, \tag{42}$$

where

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_T \otimes \mathbf{\Sigma}).$$

It then follows that

$$\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\mathbf{x}\boldsymbol{\beta}, \mathbf{I}_T \otimes \boldsymbol{\Sigma}\right),$$

and the likelihood function is given by

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= |2\pi(\mathbf{I}_T \otimes \boldsymbol{\Sigma})|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma})^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})\right\} \\ &= (2\pi)^{-\frac{Tg}{2}} |\mathbf{I}_T \otimes \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})\right\}, \end{aligned} \tag{43}$$

where the second equality holds because $|\mathbf{I}_T \otimes \boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}|^T$ and $(\mathbf{I}_T \otimes \boldsymbol{\Sigma})^{-1} = \mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1}$. We could also write the log likelihood as

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{Tg}{2}} |\mathbf{I}_T \otimes \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{x}_t\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}_t - \mathbf{x}_t\boldsymbol{\beta})\right\}.$$

So we now have an expression for the likelihood function associated with the observed data. We can then apply Bayes' Rule where, in general, we can denote $f(\mathbf{y}|\boldsymbol{\theta})$ as the observed likelihood, $f(\boldsymbol{\theta})$ as the prior distribution, and $f(\boldsymbol{\theta}|\mathbf{y})$ as the posterior distribution. Thus we have, as usual:

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{y}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{y})} \\ &\propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}). \end{aligned}$$

4.3 The Minnesota prior

The simplest form of prior distributions for VAR models is known as the Minnesota (or Litterman) prior, originally proposed by Litterman (1980, 1986). The basic intuition behind this prior is that the behaviour of most macroeconomic variables is well approximated by a random walk with drift. Hence, it centres the distribution of the coefficients in $\boldsymbol{\beta}$ at a value that implies a random-walk behaviour for all the elements in, say, y_t:

$$y_t = c + y_{t-1} + u_t.$$

While not motivated by economic theory, these are computationally convenient priors, meant to capture commonly held beliefs about how economic time series behave.

The Minnesota prior assumes that the coefficients $\mathbf{A}_1, ..., \mathbf{A}_p$ in (24) are a priori independent and

normally distributed, with the following moments:

$$\mathbb{E}\left[a_{ij}^{(l)}|\mathbf{\Sigma}\right] = \begin{cases} \delta_i, & i = j, l = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{44}$$

$$\text{Var}\left(a_{ij}^{(l)}|\mathbf{\Sigma}\right) = \begin{cases} \frac{\lambda_1^2}{f(l)}, & \text{for } i = j, \forall l, \\ \frac{\sigma_i^2}{\sigma_j^2}\left(\frac{\lambda_1\lambda_2}{f(l)}\right)^2, & \text{for } i \neq j, \forall l, \end{cases} \tag{45}$$

$$\text{Var}\left(c_{ij}|\mathbf{\Sigma}\right) = \sigma_i^2(\lambda_1\lambda_4)^2, \quad \forall i. \tag{46}$$

In the original Minnesota prior formulation, $\delta_i = 1,\ \forall i$, yielding

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{10} \\ a_{20} \end{bmatrix} + \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}$$
$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}$$
$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix},$$

which fits the assumption that each series is a random-walk. Note that this assumption may not be appropriate if the variables in $\mathbf{y}_t$ are strongly mean reverting. For series which are stationary, or transformed to achieve stationarity, you can centre the distributions around zero.

The Minnesota prior assumes that lags of other variables are less informative than the autoregressive lags, and that more recent lags are more informative than distant lags. This is captured by $f(l)$. A common choice for this function is a harmonic lag decay:

$$f(l) = l^{\lambda_3},$$

where the severity of the lag is regulated by the hyperparameter λ_3. The hyperparameters σ_i^2 and σ_j^2 are often fixed using sample information, e.g., from univariate regressions of each variable onto its own lags.

Importantly, λ_1 is a hyperparameter that controls the overall tightness of the random walk prior. If $\lambda_1 = 0$, the prior information dominates, and the VAR reduces to a vector of univariate models. Conversely, as $\lambda_1 \to \infty$ the prior becomes less informative, and the posterior mostly mirrors sample information. Meanwhile, λ_2 governs the variance of coefficient i to the lags of other coefficients, λ_3 governs the informativeness of autoregressive lags, and λ_4 covers the variance of exogenous variables.

To derive, start with the likelihood function. For the Minnesota prior, (42) is most convenient to

work with:

$$\begin{aligned} \mathbf{y} &= \mathbf{x}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_T) \\ \implies \mathbf{y} &\sim \mathcal{N}\left(\mathbf{x}\boldsymbol{\beta}, \boldsymbol{\Sigma} \otimes \mathbf{I}_T\right), \end{aligned}$$

and this gives the likelihood for $\mathbf{y}$ as

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{gT}{2}} |\Sigma \otimes \mathbf{I}_T|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\boldsymbol{\Sigma} \otimes \mathbf{I}_T)^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})\right\},$$

and ignoring terms independent of $\boldsymbol{\beta}$ gives us:

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\boldsymbol{\Sigma} \otimes \mathbf{I}_T)^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})\right\}. \tag{47}$$

Assuming that $\boldsymbol{\Sigma}$ is known,[7] the prior for $\boldsymbol{\beta}$ is given by (73), and from Bayes' Rule we have $f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$. Thus, we can simplify by keeping the kernel of the multivariate normal distribution,

$$f(\boldsymbol{\beta}) \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0\right), \tag{48}$$

and write

$$f(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\},$$

where remember that the elements of the variance-covariance matrix, $\boldsymbol{\Omega}_0$, are given by (45) and (46). The posterior for $\boldsymbol{\beta}$ is then multivariate normal, given by combining (43) and our prior distribution:

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\beta})f(\boldsymbol{\beta}) \\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})\right\} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right]\right\}, \end{aligned} \tag{49}$$

and as before, let's manipulate the term inside the square brackets. Expand it out by "completing the

[7] Or replaced with an estimate, $\hat{\boldsymbol{\Sigma}}$.

squares":

$$\begin{aligned}
&\mathbf{y}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} - \mathbf{y}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{x}\boldsymbol{\beta} - (\mathbf{x}\boldsymbol{\beta})^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} + (\mathbf{x}\boldsymbol{\beta})^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{x}\boldsymbol{\beta} \\
&+ \boldsymbol{\beta}^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 \\
=&\mathbf{y}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{x}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} + (\mathbf{x}\boldsymbol{\beta})^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{x}\boldsymbol{\beta} \\
&+ \boldsymbol{\beta}^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 \\
=&\mathbf{y}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} - 2\boldsymbol{\beta}^\top\left(\mathbf{x}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} + \boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0\right) \\
&+ \boldsymbol{\beta}^\top\left(\mathbf{x}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}\right) + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0.
\end{aligned}$$

Then, we define

$$\begin{aligned}
\bar{\boldsymbol{\beta}} &= \bar{\boldsymbol{\Omega}}\left(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{x}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y}\right) \\
&= \bar{\boldsymbol{\Omega}}\left(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}^\top)\mathbf{y}\right), \\
\bar{\boldsymbol{\Omega}} &= \left(\boldsymbol{\Omega}_0^{-1} + \mathbf{x}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{x}\right)^{-1} \\
&= \left(\boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}^\top\mathbf{X}\right)^{-1},
\end{aligned}$$

and then do our usual "add and subtract" trick:

$$\begin{aligned}
&\mathbf{y}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} - 2\boldsymbol{\beta}^\top\underbrace{\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\Omega}}}_{=\mathbf{I}}\left(\mathbf{x}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} + \boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0\right) \\
&+ \boldsymbol{\beta}^\top\left(\mathbf{x}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}\right) + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0\underbrace{+\bar{\boldsymbol{\beta}}^\top\bar{\boldsymbol{\Omega}}\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^\top\bar{\boldsymbol{\Omega}}\bar{\boldsymbol{\beta}}}_{=0},
\end{aligned}$$

and then clean up:

$$\begin{aligned}
&\mathbf{y}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} - 2\boldsymbol{\beta}^\top\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\beta}} + \boldsymbol{\beta}^\top\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + \bar{\boldsymbol{\beta}}^\top\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^\top\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\beta}} \\
=&\left[\boldsymbol{\beta}^\top\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\beta} + \bar{\boldsymbol{\beta}}^\top\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\beta}} - 2\boldsymbol{\beta}^\top\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\beta}}\right] + \mathbf{y}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}^\top\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\beta}} \\
=&(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top\bar{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + \mathbf{y}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}^\top\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\beta}}.
\end{aligned}$$

Thus, we can write (49) as

$$\begin{aligned}
f(\boldsymbol{\beta}|\mathbf{y}) &\propto \exp\left\{-\frac{1}{2}\left[(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top\boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top\bar{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right]\right\}\exp\left\{-\frac{1}{2}\left[\mathbf{y}^\top(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y} + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}^\top\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\beta}}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top\bar{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right]\right\}. \qquad (50)
\end{aligned}$$

This of course implies:

$$f(\boldsymbol{\beta}|\mathbf{y}) \sim \mathcal{N}\left(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\Omega}}\right).$$

Here the posterior for $\boldsymbol{\beta}$ is multivariate normal, much like its conjugate prior. It's once again worth noting that this is the case where $\boldsymbol{\Sigma}$ is assumed to be a known quantity, and since we're only interested in the distribution of $\boldsymbol{\beta}$, we can drop all terms not containing $\boldsymbol{\beta}$ to attain the proportional density as expressed in (50).

4.3.1 Minnesota prior example

Consider a VAR model with two endogenous variables ($g = 2$) and two lags ($p = 2$), along with one exogenous variable ($m = 1$). Each equation will involve $k = gp + m = 2 \times 2 + 1 = 5$ coefficients to estimate, implies a total of $q = nk = 2 \times 5 = 10$ coefficients for the whole model, so that $\boldsymbol{\beta}_0$ will be a $q \times 1$ vector:

$$\boldsymbol{\beta}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

For the variance covariance matrix $\boldsymbol{\Omega}_0$, it is assumed that no covariance exists between terms in $\boldsymbol{\beta}$, so that $\boldsymbol{\Omega}_0$ is a diagonal matrix. Also, Litterman (1986) argued that the further the lag, the more confident we should be that coefficients linked to this lag have a value of zero. Therefore, variances should be smaller on further lags.

- For parameters in $\boldsymbol{\beta}$ relating endogenous variables to their own lags, the variance is given by:

$$\sigma^2_{a_{ii}} = \left(\frac{\lambda_1}{l^{\lambda_3}}\right)^2,$$

  where λ_1 is the overall tightness parameter, l is the lag considered by the coefficient, and λ_3 is a scaling coefficient controlling the speed at which coefficients for lags greater than 1 converge to 0 with greater certainty.

- For parameters related to cross-variable lag coefficients, the variance is given by:

$$\sigma^2_{a_{ij}} = \frac{\sigma_i^2}{\sigma_j^2}\left(\frac{\lambda_1\lambda_2}{l^{\lambda_3}}\right)^2,$$

  where σ_i^2 and σ_j^2 denote the OLS residual variance of the autoregressive models estimated for variables i and j, and λ_2 represents a cross-variable specific variance parameter.

- For exogenous variables (including constant terms), the variance is given by:

$$\sigma^2_{c_i} = \sigma_i^2(\lambda_1\lambda_4)^2,$$

  where σ_i^2 is again the OLS residual variance of an autoregressive model previously estimated for variable i, and λ_4 is a large (potentially infinite) variance parameter.

Thus, $\mathbf{\Omega}_0$ is a $q \times q$ diagonal matrix with three different types of variance terms on its main diagonal. For instance, for the VAR model with $g = 2$, $p = 2$, and $m = 1$, we have

$$\mathbf{\Omega}_0 = \begin{bmatrix}
\lambda_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{\sigma_1^2}{\sigma_2^2}(\lambda_1\lambda_2)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{\sigma_1^2}{\sigma_2^2}\left(\frac{\lambda_1\lambda_2}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_1^2(\lambda_1\lambda_4)^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{\sigma_2^2}{\sigma_1^2}(\lambda_1\lambda_2)^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_1^2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\sigma_2^2}{\sigma_1^2}\left(\frac{\lambda_1\lambda_2}{2^{\lambda_3}}\right)^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_2^2(\lambda_1\lambda_4)^2
\end{bmatrix}$$

Different choices are possible $\lambda_1, \lambda_2, \lambda_3$, and λ_4. However, values typically found in the literature revolve around:

$$\begin{aligned}
\lambda_1 &= 0.1, \\
\lambda_2 &= 0.5, \\
\lambda_3 &= \{1, 2\}, \\
\lambda_4 &= \{10^2, ..., \infty\}.
\end{aligned}$$

Finally, since the Minnesota prior assumes that the variance covariance matrix of residuals $\mathbf{\Sigma}$ is known, one has to decide how to define it. The original Minnesota prior assumes that $\mathbf{\Sigma}$ is diagonal, which

as we've seen, conveniently implies independence between the VAR coefficients of different equations. This property was useful at a time of limited computational power as it allows estimating the model equation by equation. A first possibility is thus to set the diagonal of $\mathbf{\Sigma}$ equal to the residual variance of individual autoregressive models run on each variable in the VAR. A second possibility is to use the variance covariance matrix of a conventional VAR estimated by OLS/ML, but to retain only the diagonal terms as $\mathbf{\Sigma}$. Finally, as the model estimates all the equations simultaneously in this setting, the assumption of a diagonal matrix is not required. Therefore, a third and last possibility consists in using directly the entire variance covariance matrix of a VAR estimated by OLS/ML.

4.4 Natural conjugate Normal-Inverse-Wishart priors

A drawback of the Minnesota prior is its treatment of $\mathbf{\Sigma}$ – ideally, we want to treat $\mathbf{\Sigma}$ as an unknown parameter. The natural conjugate prior allows us to do this in a way that yields analytical results. But, as we shall see, this has some drawbacks.

The Normal-Inverse-Wishart (NIW) conjugate priors, or natural conjugate priors, are part of the exponential family and are commonly used prior distributions for $(\mathbf{\Pi}, \mathbf{\Sigma})$ in VARs with Gaussian errors. These assume a multivariate normal distribution for the regression coefficients, and an Inverse-Wishart specification for the variance covariance matrix of the error term, and can be written as

$$\boldsymbol{\beta}|\mathbf{\Sigma} \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \mathbf{\Sigma} \otimes \mathbf{\Phi}_0\right), \tag{51}$$

$$\mathbf{\Sigma} \sim \mathcal{W}^{-1}\left(\mathbf{\Psi}_0, \nu_0\right), \tag{52}$$

where $(\boldsymbol{\beta}_0, \mathbf{\Phi}_0, \mathbf{\Psi}_0, \nu_0)$ are the priors' hyperparameters and $\mathcal{W}^{-1}(\cdot)$ denotes the Inverse-Wishart distribution. Similar to the Minnesota prior, $\boldsymbol{\beta}_0$ is a $q \times 1$ vector, $\mathbf{\Phi}_0$ is a $k \times k$ diagonal matrix, and $\mathbf{\Sigma}$ is the usual VAR residual variance covariance matrix, which implies that $\mathbf{\Sigma} \otimes \mathbf{\Phi}_0$ is a $gk \times gk$ or $q \times q$ variance covariance matrix.

Definition 4 (**Inverse-Wishart Distribution**)**.** An $m \times m$ random matrix $\mathbf{Z}$ is said to have an Inverse-Wishart distribution, $\mathbf{Z} \sim \mathcal{W}^{-1}(\mathbf{\Psi}, \nu)$, with shape parameter $\nu > 0$ and scale matrix $\mathbf{\Psi}$ if its density function is given by

$$f(\mathbf{Z}; \mathbf{\Psi}, \nu) = \frac{|\mathbf{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{m\nu}{2}} \Gamma_m\left(\frac{\nu}{2}\right)} |\mathbf{Z}|^{-\frac{\nu+m+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Psi}\mathbf{Z}^{-1}\right)\right\},$$

where $\Gamma_m(\cdot)$ is the multivariate gamma function and $\mathrm{tr}(\cdot)$ is the trace function. For $\nu > m + 1$, we have

$$\mathbb{E}[\mathbf{Z}] = \frac{\mathbf{\Psi}}{\nu - m - 1}.$$

To derive, begin with the likelihood function given by Equation (43) again,

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{Tg}{2}} |\mathbf{I}_T \otimes \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})\right\}.$$

But now since $\boldsymbol{\Sigma}$ is assumed to be unknown, $\mathbf{I}_T \otimes \boldsymbol{\Sigma}$ cannot be disregarded as a proportionality constant. Thus, we can can only simplify down to

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\mathbf{I}_T \otimes \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})\right\}.$$

After some algebraic manipulation, one can write this density as:

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{(p+1)g}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)^\top \left(\boldsymbol{\Sigma} \otimes (\mathbf{X}^\top \mathbf{X})^{-1}\right)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \\ \times |\boldsymbol{\Sigma}|^{-\frac{T-(p+1)g}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Pi}})^\top (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Pi}})\right)\right\}. \end{aligned} \tag{53}$$

Choosing $\boldsymbol{\beta}_0$ is quite simple, but the choice of $\boldsymbol{\Phi}_0$ is quite difficult. We can adopt a Minnesota prior scheme for $\boldsymbol{\beta}_0$, setting values around 1 for each element's first-lag components, and 0 for cross variable and exogenous coefficients. For $\boldsymbol{\Phi}_0$ note the difference between the Minnesota prior (48) and the natural conjugate prior (51): while $\boldsymbol{\Omega}_0$ represents the full variance covariance matrix of $\boldsymbol{\beta}$, $\boldsymbol{\Phi}_0$ only represents the variance for the parameters of one single equation in the VAR. Each such variance is then scaled by the variable specific variance in $\boldsymbol{\Sigma}$. This Kronecker structure implies that the variance covariance matrix of one equation has to now be proportional to the variance covariance matrix of the other equations, unlike in $\boldsymbol{\Omega}_0$ in the Minnesota prior.

We could proceed as follows however: for lag terms (both own and cross lags), define the variance as

$$\sigma^2_{a_{ij}} = \frac{1}{\sigma^2_j}\left(\frac{\lambda_1}{l^{\lambda_3}}\right)^2,$$

where σ^2_j is the unknown residual variance for variable j in the BVAR model, approximated by individual AR regressions. For exogenous variables, define the variance as:

$$\sigma^2_c = (\lambda_1 \lambda_4)^2.$$

For instance with the $g = 2$, $p = 2$, and $m = 1$ VAR model, $\mathbf{\Phi}_0$ would be

$$\mathbf{\Phi}_0 = \begin{bmatrix} \frac{1}{\sigma_1^2}\lambda_1^2 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma_2^2}\lambda_1^2 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sigma_1^2}\left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sigma_2^2}\left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & (\lambda_1\lambda_4)^2 \end{bmatrix},$$

and if one assumes a diagonal $\mathbf{\Sigma}$ as in the original Minnesota prior, we would have

$$\mathbf{\Sigma}\otimes\mathbf{\Phi}_0 = \begin{bmatrix} \lambda_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\sigma_1^2}{\sigma_2^2}(\lambda_1\lambda_2)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\sigma_1^2}{\sigma_2^2}\left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_1^2(\lambda_1\lambda_4)^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\sigma_2^2}{\sigma_1^2}\lambda_1^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\sigma_2^2}{\sigma_1^2}\left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_2^2(\lambda_1\lambda_4)^2 \end{bmatrix}.$$

If we compare this with the expression for $\mathbf{\Omega}_0$ in the Minnesota prior, we see that $\mathbf{\Sigma} \otimes \mathbf{\Phi}_0$ is a special case of $\mathbf{\Omega}_0$ when $\lambda_2 = 1$. In this sense, the natural conjugate NIW prior appears as a Minnesota prior that would not be able to provide tighter priors on cross-variable parameters, which may be an undesirable assumption. For this reason, it is advised to set λ_1 at a smaller value than for the Minnesota prior (e.g., between 0.01 and 0.1), in order to compensate for the lack of extra shrinkage from λ_2. For the remaining hyperparameters, λ_3 and λ_4, the same values as the Minnesota prior may be applied.

With $\boldsymbol{\beta}_0$ and $\mathbf{\Phi}_0$ in hand, the prior density for $\boldsymbol{\beta}$ can be written as:

$$f(\boldsymbol{\beta}) \propto |\mathbf{\Sigma}|^{-\frac{gp+m}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^\top \left(\mathbf{\Sigma}\otimes\mathbf{\Phi}_0\right)^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right\}. \tag{54}$$

As for $\mathbf{\Sigma}$, we have:

$$\mathbf{\Sigma} \sim \mathcal{W}^{-1}\left(\mathbf{\Psi}_0, \nu_0\right),$$

where $\mathbf{\Psi}_0$ is a $g \times g$ scale matrix for the prior, and ν_0 is prior degrees of freedom. While any choice can be made for these hyperparameters according to prior information, the literature once again proposes

standard schemes. For example, $\mathbf{\Psi}_0$ can be

$$\mathbf{\Psi}_0 = (\nu_0 - g - 1) \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_g^2 \end{bmatrix},$$

and

$$\nu_0 = g + 2.$$

This specifies the prior degrees of freedom as the minimum possible to obtain well-defined mean and variance. Indeed, this value implies that

$$\mathbb{E}\left[\mathbf{\Sigma}\right] = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_g^2 \end{bmatrix}.$$

In other words, the expectation of $\mathbf{\Sigma}$ is the diagonal variance covariance matrix obtained from individual AR regressions and used as an estimate for $\mathbf{\Sigma}$ in the Minnesota prior. As with the Minnesota prior, it is possible to implement alternative schemes.

With these elements, the kernel of the prior density for $\mathbf{\Sigma}$ is given by:

$$f(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\frac{\nu_0+g+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Psi}_0\right)\right\}. \tag{55}$$

Combine the likelihood and the priors to get

$$\begin{aligned} f(\boldsymbol{\beta}, \mathbf{\Sigma}|\mathbf{y}) \propto |\Sigma|^{-\frac{gp+m}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\right)^\top \left(\mathbf{\Sigma} \otimes \bar{\mathbf{\Phi}}\right)^{-1} \left(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\right)\right\} \\ \times |\mathbf{\Sigma}|^{-\frac{\bar{\nu}+g+1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{\Sigma}^{-1}\bar{\mathbf{\Psi}}\right)\right\}, \end{aligned} \tag{56}$$

where

$$\boldsymbol{\beta}|\mathbf{\Sigma}, \mathbf{y} \sim \mathcal{N}\left(\bar{\boldsymbol{\beta}}, \mathbf{\Sigma} \otimes \bar{\mathbf{\Phi}}\right), \tag{57}$$

$$\mathbf{\Sigma}|\mathbf{y} \sim \mathcal{W}^{-1}\left(\bar{\mathbf{\Psi}}, \bar{\nu}\right), \tag{58}$$

which is why refer to this as the "natural conjugate Normal-Inverse-Wishart" prior distribution, and

where we also have

$$\begin{aligned}
\bar{\boldsymbol{\beta}} &= \text{vec}\left(\bar{\boldsymbol{\Pi}}\right) \\
&= \text{vec}\left(\bar{\boldsymbol{\Phi}}\left[\boldsymbol{\Phi}_0^{-1}\boldsymbol{\Pi}_0 + \mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\Pi}}\right]\right), \qquad (59)\\
\bar{\boldsymbol{\Phi}} &= \left(\boldsymbol{\Phi}_0 + \mathbf{X}^\top\mathbf{X}\right)^{-1}, \qquad (60)\\
\bar{\boldsymbol{\Psi}} &= \hat{\boldsymbol{\Pi}}^\top\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\Pi}} + \boldsymbol{\Pi}_0^\top\boldsymbol{\Phi}_0^{-1}\boldsymbol{\Pi}_0 + \hat{\mathbf{e}}^\top\hat{\mathbf{e}} - \bar{\boldsymbol{\Pi}}^\top(\boldsymbol{\Phi}^{-1} + \mathbf{X}^\top\mathbf{X})\bar{\boldsymbol{\Pi}}, \qquad (61)\\
\bar{\nu} &= T + \nu_0. \qquad (62)
\end{aligned}$$

Note that obtaining the marginal distribution for $\boldsymbol{\Sigma}$ from the posterior (56) is trivial: integrating out $\boldsymbol{\beta}$ is easy as it appears only in the first term as a multivariate normal. Following integration, only the second term remains, which is the kernel of an inverse Wishart distribution, and thus we get (58).

We can then integrate out $\boldsymbol{\Sigma}$ to derive the marginal distribution for $\boldsymbol{\Pi}$:

$$f(\boldsymbol{\Pi}|\mathbf{y}) \propto \left|\mathbf{I}_g + \bar{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\Pi} - \bar{\boldsymbol{\Pi}})^\top\bar{\boldsymbol{\Phi}}^{-1}(\boldsymbol{\Pi} - \bar{\boldsymbol{\Pi}})\right|^{-\frac{T+\nu_0-g+1+g+gp+m-1}{2}}. \qquad (63)$$

This is the kernel of a matrix student distribution with mean $\bar{\boldsymbol{\Pi}}$, scale matrices $\bar{\boldsymbol{\Psi}}$ and $\bar{\boldsymbol{\Phi}}$, and degrees of freedom $T + \nu_0 - g + 1$:

$$\boldsymbol{\Pi} \sim \mathcal{T}\left(\bar{\boldsymbol{\Pi}}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\Phi}}, \tilde{\nu}\right),$$

with

$$\tilde{\nu} = T + \nu_0 - g + 1.$$

This then implies that each individual element $\Pi_{i,j}$ of $\boldsymbol{\Pi}$ follows a univariate student distribution with mean $\bar{\Pi}_{i,j}$, scale parameter $\bar{\Phi}_{i,j} \times \bar{\Psi}_{j,j}$, and degrees of freedom $\tilde{\nu}$,

$$\Pi_{i.j} \sim t(\bar{\Pi}_{i,j}, \bar{\Phi}_{i,i} \times \bar{\Psi}_{j,j}, \tilde{\nu}).$$

These statistics can be used to compute point estimates and draw inference for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

4.4.1 Natural conjugate Normal-Inverse-Wishart prior as dummy observables

The informative NIW priors in Equations (51)-(52) can be thought of as equivalent to having observed dummy "pseudo" observations $(\mathbf{Y}^d, \mathbf{X}^d)$ of size T^d such that

$$\begin{aligned}
\boldsymbol{\Psi}_0 &= (\mathbf{Y}^d - \mathbf{X}^d\boldsymbol{\Pi}_0)^\top(\mathbf{Y}^d - \mathbf{X}^d\boldsymbol{\Pi}_0), \\
\nu_0 &= T^d - gp - m, \\
\boldsymbol{\beta}_0 &= \text{vec}(\boldsymbol{\Pi}_0) \\
&= \text{vec}\left(\left[(\mathbf{X}^d)^\top\mathbf{X}^d\right]^{-1}(\mathbf{X}^d)^\top\mathbf{Y}^d\right), \\
\boldsymbol{\Phi}_0 &= \left[(\mathbf{X}^d)^\top\mathbf{X}^d\right]^{-1}.
\end{aligned}$$

Once a set of pseudo-observations able to match the wished hyperparameters is found, the posterior can be equivalently estimated using the extended samples:

$$\begin{aligned}
\mathbf{Y}^* &= \begin{bmatrix}\mathbf{Y} \\ \mathbf{Y}^d\end{bmatrix}, \\
\mathbf{X}^* &= \begin{bmatrix}\mathbf{X} \\ \mathbf{X}^d\end{bmatrix},
\end{aligned}$$

which are of size $T^* = T + T^d$. We can then obtain:

$$\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{\Sigma}, \mathbf{y} &\sim \mathcal{N}\left(\boldsymbol{\beta}^*, \boldsymbol{\Sigma} \otimes \left[(\mathbf{X}^*)^\top\mathbf{X}\right]^{-1}\right), \\
\boldsymbol{\Sigma}|\mathbf{y} &\sim \mathcal{W}^{-1}\left(\boldsymbol{\Psi}^*, T^* + \nu_0\right).
\end{aligned}$$

Usually, it's simply not possible to sample directly from the posterior distribution. As such, MCMC algorithms are used.

The natural conjugate NIW prior has great advantages because we're able to yield analytical results. Also, notice that in Equations (51)-(52) we have a Kronecker product in the Inverse-Wishart distribution's scale factor. This comes from the definition of $\mathbf{x}_t = (\mathbf{I}_g \otimes \mathbf{X}_t)$, which means that every equation must have the same set of explanatory variables. As such, because the same set of regressors appear in each equation, homoskedastic VARs can be written as SUR models. This symmetry across equations means that homoskedastic VAR models have a Kronecker product in the likelihood, which in turn implies that estimation can be broken into g separate least-squares calculations, each only of dimension $gp + m$. The symmetry in the likelihood can be inherited by the posterior, if the prior adopted also features a Kronecker product as in Equation (51).

But this has problems which make it rarely used in practice. Consider the following example: a VAR which involves variables such as output growth and the growth in the money supply, where the

researcher wants to impose the neutrality of money assumption. This implies that the coefficients on the lagged money growth variables in the output growth equation are zero (but coefficients of lagged money growth in other equations would not be zero). But as we just discussed, the $\mathbf{x}_t = \mathbf{I}_g \otimes \mathbf{X}_t$ form of the explanatory variables means that every equation must have same set of explanatory variables. But if we do not maintain the $\mathbf{x}_t$ form, then we don't get analytical conjugate prior.

The other problem is that we cannot "almost impose" a neutrality of money restriction through the prior – i.e., we cannot set prior mean over neutrality of money restriction and set prior variance to a very small value. To see why, let individual elements of $\boldsymbol{\Sigma}$ be σ_{ij}, and so the prior covariance matrix has form $\boldsymbol{\Sigma} \otimes \boldsymbol{\Phi}_0$. This implies prior covariance of coefficients in equation i is $\sigma_{ii}\boldsymbol{\Phi}_0$. Thus, prior covariance of the coefficients in any two equations must be proportional to one another. So while we can "almost impose" coefficients on lagged money growth to be zero in all equations, we cannot do it in a single equation![8]

4.5 Independent priors

Up until now, we haven't really used much economic theory – it's just been brute forced algebra regarding matrices and VAR models – and we haven't departed too far from standard econometric methods for VAR models. But let's stop and consider Bayes' Rule again. When we looked at linear regression models, we had to make some strong assumptions about the prior density $f(\boldsymbol{\theta})$ when we wanted to estimate $\boldsymbol{\beta}$ and σ^2. Recall assumption (19) about the prior densities for $\boldsymbol{\beta}$ and σ^2, $f(\boldsymbol{\beta})$ and $f(\sigma^2)$, being independent which allowed us to write $f(\boldsymbol{\theta})$ as the product of $f(\boldsymbol{\beta})$ and $f(\sigma^2)$.

Here, for the VAR case, we're going to make a similar assumption, namely,

$$f(\boldsymbol{\theta}) = f(\boldsymbol{\beta})f(\boldsymbol{\Sigma}). \tag{64}$$

Below we cover some common assumptions regarding the prior distributions, $f(\boldsymbol{\beta})$ and $f(\boldsymbol{\Sigma})$.

4.5.1 Diffuse (Jeffreys') prior

A possible alternative to the Minnesota and natural conjugate NIW priors is the normal-diffuse, or Jeffreys', prior. It gets its namesake because it relies on a diffuse (uninformative) prior for $\boldsymbol{\Sigma}$. The Jeffreys' prior is a good alternative to the independent NIW prior in Section 4.5.2 when one wants to remain agnostic about the value that $\boldsymbol{\Sigma}$ should be given. The likelihood function and the prior distribution for $\boldsymbol{\beta}$ are similar to those developed in the previous subsections and are thus respectively

[8] Note that the Minnesota prior form for $\boldsymbol{\Phi}_0$ is not consistent with natural conjugate prior.

given by:

$$f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-T/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^{\top}\left[\boldsymbol{\Sigma}\otimes(\mathbf{X}^{\top}\mathbf{X})^{-1}\right]\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right\}$$
$$\times \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})^{\top}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})\right)\right\}, \tag{65}$$

and

$$f(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^{\top}\boldsymbol{\Omega}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right\}. \tag{66}$$

The main focal point of the diffuse prior is the prior distribution for $\boldsymbol{\Sigma}$, which is:

$$f(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(g+1)/2}. \tag{67}$$

This prior is called an improper prior as it integrates to infinity rather than to one. Yet, this does not necessarily preclude the posterior distribution to be proper, which is indeed the case here.

Jeffreys' priors are proportional to the square root of the determinant of the Fisher information matrix, and are derived from the Jeffreys' "invariance principle", meaning that the prior is invariant to re-parameterisation. Essentially, the priors act as non-informative or flat priors, and are designed to extract the maximum amount of expected information from the data. They maximise the difference (measured by the Kullback-Leibler distance) between the posterior and the prior when the number of samples drawn goes to infinity.

With these priors, we can write the posterior distribution of the VAR parameters as

$$\begin{aligned} f(\boldsymbol{\beta},\boldsymbol{\Sigma}|\mathbf{y}) &\propto |\boldsymbol{\Sigma}|^{-T/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^{\top}\left[\boldsymbol{\Sigma}\otimes(\mathbf{X}^{\top}\mathbf{X})^{-1}\right]\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right\} \\ &\times \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})^{\top}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})\right)\right\} \\ &\times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^{\top}\boldsymbol{\Omega}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right\} \\ &\times |\boldsymbol{\Sigma}|^{-(g+1)/2} \\ &\propto |\boldsymbol{\Sigma}|^{-\frac{T+g+1}{2}} \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^{\top}\left[\boldsymbol{\Sigma}\otimes(\mathbf{X}^{\top}\mathbf{X})^{-1}\right]\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)+\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^{\top}\boldsymbol{\Omega}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right]\right\} \\ &\times \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})^{\top}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})\right)\right\}. \end{aligned} \tag{68}$$

This can be further simplified to

$$f(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}) \propto |\boldsymbol{\Sigma}|^{-\frac{T+g+1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)^{\top}\bar{\boldsymbol{\Omega}}^{-1}\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)\right\}$$
$$\times \exp\left\{-\frac{1}{2}\left[\hat{\boldsymbol{\beta}}^{\top}\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}^{\top}\mathbf{X}\right)\hat{\boldsymbol{\beta}}+\boldsymbol{\beta}_0^{\top}\boldsymbol{\Omega}_0\boldsymbol{\beta}_0-\bar{\boldsymbol{\beta}}^{\top}\bar{\boldsymbol{\Omega}}\bar{\boldsymbol{\beta}}\right]\right\}$$
$$\times \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})^{\top}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})\right)\right\},$$

where

$$\bar{\boldsymbol{\beta}} = \bar{\boldsymbol{\Omega}}\left[\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + (\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}^{\top})\mathbf{y}\right], \tag{69}$$
$$\bar{\boldsymbol{\Omega}} = \left(\boldsymbol{\Omega}_0^{-1}+\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}^{\top}\mathbf{X}\right)^{-1}. \tag{70}$$

The posterior distribution for $\boldsymbol{\beta}$ is obtained by ignoring any term not involving $\boldsymbol{\beta}$:

$$f(\boldsymbol{\beta}|\boldsymbol{\Sigma}, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)^{\top}\bar{\Omega}^{-1}\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)\right\}, \tag{71}$$

which is recognised as the kernel of a multivariate normal distribution:

$$f(\boldsymbol{\beta}|\boldsymbol{\Sigma}, \mathbf{y}) \sim \mathcal{N}\left(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\Omega}}\right).$$

Meanwhile, the conditional posterior distribution for $\boldsymbol{\Sigma}$ is obtained from (68) by ignoring proportional constants not involving $\boldsymbol{\Sigma}$ and rearranging:

$$f(\boldsymbol{\Sigma}|\boldsymbol{\beta}, \mathbf{y}) \propto |\boldsymbol{\Sigma}|^{-\frac{T+g+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left[(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})^{\top}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})\right]\right)\right\}. \tag{72}$$

This is the kernel of an Inverse-Wishart distribution:

$$f(\boldsymbol{\Sigma}|\boldsymbol{\beta}, \mathbf{y}) \sim \mathcal{W}^{-1}\left(\bar{\boldsymbol{\Psi}}, T-(p+1)g\right),$$

where

$$\bar{\boldsymbol{\Psi}} = (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})^{\top}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}}).$$

Given the diffuse priors on $\boldsymbol{\Sigma}$, the posterior for the autoregressive coefficients is centred at the MLE, with posterior variance $\boldsymbol{\Sigma}\otimes(\mathbf{X}^{\top}\mathbf{X})^{-1}$. Under these assumptions, Bayesian probability statements about the parameters (given the data) have the same form as the frequentist pre-sample probability statements about the parameter estimators. In general, under widely applicable regularity conditions

for a given estimator, $\hat{\boldsymbol{\beta}}^*$, where

$$\sqrt{T}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})|\boldsymbol{\beta} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

one can approximate the distribution $\sqrt{T}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^*)|\hat{\boldsymbol{\beta}}$ as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ in large samples. Hence, we can interpret $(1-\alpha)$ approximate confidence regions generated from the frequentist asymptotic approximate distribution as if they were sets in the parameter space with posterior probability $(1-\alpha)$.

But, one set of assumptions in which there is a significant divergence between pre-sample frequentist probability statements and Bayesian post-sample probability statements is the case of time series models with unit roots. In such cases, while the frequentist distribution of the estimator is skewed asymptotically, the posterior density function is unaffected.

4.5.2 Independent Normal-Inverse-Wishart priors

The natural conjugate NIW prior had $\boldsymbol{\beta}|\boldsymbol{\Sigma}$ being normal and $\boldsymbol{\Sigma}$ being Inverse-Wishart, and the VAR had the same explanatory variables in every equation. There, assuming an unknown $\boldsymbol{\Sigma}$ comes at the cost of imposing a Kronecker structure on the prior distribution for $\boldsymbol{\beta}$, constraining its variance covariance matrix to be equal to $\boldsymbol{\Sigma} \otimes \boldsymbol{\Phi}_0$. This structure creates, for each equation, a dependence between the variance of the residual term and the variance of the VAR coefficients, which may be an undesirable assumption. We want a more general setup without these restrictive features.

We can do this with a prior for the VAR coefficients and $\boldsymbol{\Sigma}$ being independent (hence the name "independent NIW prior"). The independent Normal-Inverse-Wishart priors are commonly used prior distributions for $(\boldsymbol{\Pi}, \boldsymbol{\Sigma})$ in VAR models with Gaussian errors. These assume a multivariate normal distribution for the regression coefficients, and an Inverse-Wishart distribution for the variance-covariance matrix of the error term. An alternative way to see the restrictions generated by the natural conjugate NIW prior is to notice that the variance covariance matrix for the VAR coefficients $\boldsymbol{\Sigma} \otimes \boldsymbol{\Phi}_0$ correspond to the more general variance covariance matrix used for the Minnesota prior in the special case where $\lambda_2 = 1$. That is, where the variance on cross-variable coefficients is as large as the variance on its own lags for each equation.

Ideally, we would like to estimate BVAR model with $\boldsymbol{\Sigma}$ being treated as unknown, and an arbitrary structure could be proposed for $\boldsymbol{\Omega}_0$, with no assumed dependence between residual variance and coefficient variance. Such a prior, known as the independent Normal-Inverse-Wishart prior, is feasible but implies the sacrifice of analytical solutions in favour of numerical methods. The analysis starts the usual way: first obtain the likelihood from the data. There is no change here and the likelihood is given by:

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-T/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^\top \left[\boldsymbol{\Sigma} \otimes (\mathbf{X}^\top \mathbf{X})^{-1}\right]^{-1} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right\}$$
$$\times \exp\left\{-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\hat{\Pi})^\top(\mathbf{Y} - \mathbf{X}\hat{\Pi})\right)\right\}.$$

For the VAR(p) system with parameters $\mathbf{\Pi}$ (or, equivalently, $\boldsymbol{\beta}$) and $\mathbf{\Sigma}$, our independent priors are:[9]

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{\Omega}_0), \tag{73}$$

$$\mathbf{\Sigma} \sim \mathcal{W}^{-1}(\mathbf{\Psi}_0, \nu_0), \tag{74}$$

where $(\boldsymbol{\beta}_0, \mathbf{\Omega}_0, \mathbf{\Psi}_0, \nu_0)$ are the hyperparameters for the priors. [10] Note here here that $\mathbf{\Omega}_0$ is now an arbitrary $g^2(p+1) \times g^2(p+1)$ matrix, not necessarily adopting a Kronecker structure. $\boldsymbol{\beta}_0$, on the other hand, is the usual $g^2(p+1)$ mean vector. In typical applications, $\mathbf{\Omega}_0$ will take the form of the Minnesota variance covariance matrix, but any choice is possible. Similarly, $\boldsymbol{\beta}_0$will typically be defined as the Minnesota $\boldsymbol{\beta}_0$ vector, but any structure of vector $\boldsymbol{\beta}_0$ could be adopted.

Combining the likelihood function and our prior distributions gives us our posterior:

$$\begin{aligned} f(\boldsymbol{\beta}, \mathbf{\Sigma}|\mathbf{y}) \propto |\mathbf{\Sigma}|^{-T/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^\top \left[\mathbf{\Sigma}\otimes(\mathbf{X}^\top\mathbf{X})^{-1}\right]^{-1}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right\} \\ \times \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})^\top(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})\right)\right\} \\ \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\mathbf{\Omega}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\} \\ \times |\mathbf{\Sigma}|^{-\frac{\nu_0+g+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Psi}_0\right)\right\}, \end{aligned}$$

and doing some cleaning up yields:

$$\begin{aligned} f(\boldsymbol{\beta}, \mathbf{\Sigma}|\mathbf{y}) \propto |\mathbf{\Sigma}|^{-\frac{T+\nu_0+g+1}{2}} \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\left[\mathbf{\Sigma}^{-1}\otimes(\mathbf{X}^\top\mathbf{X})\right](\boldsymbol{\beta}-\boldsymbol{\beta}_0) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\mathbf{\Omega}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right]\right\} \\ \times \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}\left[\mathbf{\Psi}_0 + (\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})^\top(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})\right]\right)\right\}. \end{aligned}$$

[9] We are reusing our previous notation here where $\mathbf{\Omega}_0 = \text{Var}(\boldsymbol{\beta})$ is our prior for the variance covariance matrix of the coefficient vector.

[10] This prior of course implies:

$$\begin{aligned} f(\boldsymbol{\beta}) &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\mathbf{\Omega}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}, \\ f(\mathbf{\Sigma}) &\propto |\mathbf{\Sigma}|^{-\frac{\nu_0+g+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Psi}_0\right)\right\}. \end{aligned}$$

Finally, we can write this posterior density as

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}) \propto |\boldsymbol{\Sigma}|^{-\frac{T+\nu_0+g+1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^\top \bar{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\right\} \\ \times \exp\left\{-\frac{1}{2}\hat{\boldsymbol{\beta}}^\top(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}^\top\mathbf{X})\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}_0^\top\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}^\top\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\beta}}\right\} \\ \times \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{\Psi}_0 + (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})^\top(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})\right]\right)\right\}, \end{aligned} \tag{75}$$

with

$$\begin{aligned} \bar{\boldsymbol{\beta}} &= \bar{\boldsymbol{\Omega}}\left[\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + (\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}^\top)\mathbf{y}\right], \\ \bar{\boldsymbol{\Omega}} &= \left[\boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}^\top\mathbf{X}\right]^{-1}. \end{aligned}$$

As it stands, due to the general structure of Ω_0, there is no way to attain an analytical solution for the marginal distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

But, we can derive conditional distributions and a Gibbs sampler. For the VAR(p) model with likelihood function (43) and priors (73) and (74), denote the two conditional densities as: $f(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\Sigma})$ and $f(\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta})$.

The first one for $\boldsymbol{\beta}$ is easy to write out, as standard linear regression results apply:

$$\begin{aligned} &\boldsymbol{\beta}|\boldsymbol{\Sigma}, \mathbf{y} \sim \mathcal{N}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\Omega}}), \\ \implies &f(\boldsymbol{\beta}|\boldsymbol{\Sigma}, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^\top\bar{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\right\}. \end{aligned} \tag{76}$$

Next, we deal with $f(\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta})$. First, note the following "trace rule":

$$\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA}) = \text{tr}(\mathbf{CAB}).$$

Now, combine the likelihood, $f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma})$, and the prior, $f(\boldsymbol{\Sigma})$, to get

$$\begin{aligned} f(\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta}) &\propto f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) f(\boldsymbol{\Sigma}) \\ &\propto |\boldsymbol{\Sigma}|^{-\frac{T+\nu_0+g+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{\Psi}_0 + (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})^\top(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\Pi}})\right]\right)\right\}, \end{aligned} \tag{77}$$

where we have dropped proportional constants. Note that this is the kernel of an Inverse-Wishart density. In fact, we have

$$\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta} \sim \mathcal{W}^{-1}\left(\bar{\boldsymbol{\Psi}}, \nu_0 + T\right),$$

where

$$\bar{\boldsymbol{\Psi}} = \boldsymbol{\Psi}_0 + (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Pi}})^\top (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Pi}}).$$

The Gibbs sampler can then be summarised as follows:

1. Fix starting values $\boldsymbol{\beta}_{(0)}$ and $\boldsymbol{\Sigma}_{(0)}$ for two parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.
2. Draw first value of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_{(1)}$, from the conditional posterior, $f(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\Sigma}_{(0)})$ (multivariate normal as in (76)).
3. Draw first value of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{(1)}$ from the conditional posterior, $f(\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta}_{(1)})$ (Inverse-Wishart as in (77)), using $\boldsymbol{\beta}_{(1)}$.
4. Start a new cycle: draw value $\boldsymbol{\beta}_{(2)}$ from the conditional posterior, $f(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\Sigma}_{(1)})$, using $\boldsymbol{\Sigma}_{(1)}$.
5. Draw value $\boldsymbol{\Sigma}_{(2)}$ from the conditional posterior, $f(\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta}_{(2)})$, using $\boldsymbol{\beta}_{(2)}$.
6. Repeat process S times.

Note that $\boldsymbol{\Omega}_0$ can be set to anything that a researcher chooses (not restricting like the $\boldsymbol{\Sigma} \otimes \boldsymbol{\Phi}_0$ form of the natural conjugate NIW prior) – e.g., $\boldsymbol{\beta}_0$ and $\boldsymbol{\Omega}_0$ can be set as in the Minnesota prior, or setting $\nu_0 = \boldsymbol{\Psi}_0 = \boldsymbol{\Omega}_0 = 0$ gives us the uninformative prior.

4.6 Dummy observation priors

We briefly covered this in Section 4.4, but it's worth spending a bit more time on this. Most of the BVAR applications covered so far have been relying on the prior structure of the Minnesota prior. That is, for a VAR model with g endogenous variables, m exogenous variables, and p lags, the prior mean for the VAR coefficients is a $q \times 1 = g(gp + m) \times 1$ vector, $\boldsymbol{\beta}_0$, while the prior variance covariance matrix is a $q \times q$ matrix, $\boldsymbol{\Omega}_0$, with variance terms on the diagonal, and zero entries off diagonal, implying no prior covariance between the coefficients.

While this representation is convenient, it results in three main shortcomings. The first is technical and linked to the estimation of large models. Indeed, for all the priors adopting this Minnesota structure, estimation of the posterior mean, $\bar{\boldsymbol{\beta}}$, and the posterior variance, $\bar{\boldsymbol{\Omega}}$, involves the inversion of a $q \times q$ matrix. For instance, in the case of a large model with 20 endogenous variables, 20 exogenous variables, and 10 lags, the number of rows and columns of $\boldsymbol{\Omega}_0$ would be $20(20 \times 10 + 20) = 4,400$. This means that each iteration of the Gibbs sampler requires the inversion of a $4,400 \times 4,400$ matrix, rendering the process so slow that it becomes practically intractable. In the worst case, such very large matrices may even cause numerical software to fail the inversion altogether.

The second shortcoming is theoretical: with this structure, no prior covariance is assumed among the VAR coefficients, which may be suboptimal. Of course, one could simply add off-diagonal terms

in $\mathbf{\Omega}_0$ in order to create prior covariance terms. However, there is no no all-ready theory to indicate what those values should be.

The third issue is that with this kind of structure, it is very difficult to impose priors on combinations of VAR coefficients, which can yet be useful when working with unit root or co-integrated processes.

To remedy these shortcomings, in this section we will look at the dummy coefficient prior.

Consider first the prior distribution. As shown by Equation (41), it is possible to express the likelihood function for the data:

$$f(\mathbf{Y}|\mathbf{\Pi},\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})^{\top}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})\right)\right\}$$
$$\times \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{\Pi}-\hat{\mathbf{\Pi}})^{\top}\mathbf{X}^{\top}\mathbf{X}(\mathbf{\Pi}-\hat{\mathbf{\Pi}})\right)\right\},$$

where

$$\hat{\mathbf{\Pi}} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{Y}.$$

This likelihood function is then combined with a joint improper prior for $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$:

$$f(\boldsymbol{\beta},\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-(g+1)/2},$$

which is the simplest and least informative prior that one can propose for a VAR model. Combining the likelihood function with the improper prior, one obtains the posterior distribution as:

$$f(\boldsymbol{\beta},\mathbf{\Sigma}|\mathbf{y}) \propto |\mathbf{\Sigma}|^{-\frac{T+g+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})^{\top}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\Pi}})\right)\right\}$$
$$\times \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{\Pi}-\hat{\mathbf{\Pi}})^{\top}\mathbf{X}^{\top}\mathbf{X}(\mathbf{\Pi}-\hat{\mathbf{\Pi}})\right)\right\},$$

which looks like the product of a Inverse-Wishart distribution and matrix normal distribution.

A few remarks about the above posterior distribution:

1. As the product of a matrix normal distribution with an Inverse-Wishart distribution, this posterior is immediately comparable to that obtained for the NIW prior. It can then be shown that similarly to the NIW prior, the marginal posterior distributions for $\mathbf{\Sigma}$ and $\mathbf{\Pi}$ are respectively Inverse-Wishart and matrix student. They are parameterised as:

$$\mathbf{\Sigma} \sim \mathcal{W}^{-1}\left(\hat{\mathbf{\Psi}},\hat{\nu}\right),$$

with

$$\hat{\mathbf{\Psi}} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}})^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}}),$$
$$\hat{\nu} = T - (gp + m).$$

We also have

$$\mathbf{\Pi} \sim \mathcal{T}(\hat{\mathbf{\Pi}}, \hat{\mathbf{\Psi}}, \hat{\mathbf{\Phi}}, \tilde{\nu}),$$

with

$$\hat{\mathbf{\Phi}} = (\mathbf{X}^\top \mathbf{X})^{-1},$$
$$\tilde{\nu} = T - g - (gp + m) + 1.$$

2. This prior solves the dimensionality issue. While the Minnesota requires the inversion of a $q \times q$ matrix, it is apparent that this prior only requires the inversion of a $(gp+m) \times (gp+m)$ matrix. The intuition behind the result is similar to that of the natural conjugate NIW: the posterior is computed at the scale of individual equations, rather than for the full model simultaneously.

3. An uninformative prior for $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$ yields posterior estimates centred at OLS/ML values. By not providing any prior information on the mean of the estimates, and setting a flat distribution with infinite variance, one does hardly more than performing OLS estimation, using only the information provided by the data.

4. But just getting OLS estimates has a drawback: The strength of Bayesian estimation is precisely to be able to supplement the information contained in the data with personal information, in order to inflect the estimates provided by the data and improve the accuracy of the model. If one does not provide any information at all, there is, in fact, very little point in using Bayesian methods. Ideally, one would thus like to provide prior information for the model, despite the diffuse prior. We can do this with dummy observations.

Consider the possibility of generating artificial data for the model, $\mathbf{Y}^d$ and $\mathbf{X}^d$:

$$\mathbf{Y}_d = \begin{bmatrix} \text{diag}\left(\frac{\rho\sigma_1}{\lambda_1}, ..., \frac{\rho\sigma_n}{\lambda_1}\right) \\ \mathbf{O}_{g(p-1)\times g} \\ \mathbf{O}_{m\times g} \\ \text{diag}\left(\sigma_1, ..., \sigma_n\right) \end{bmatrix}, \tag{78}$$

$$\mathbf{X}_d = \begin{bmatrix} \mathbf{J}_p \otimes \text{diag}\left(\frac{\rho\sigma_1}{\lambda_1}, ..., \frac{\rho\sigma_n}{\lambda_1}\right) & \mathbf{O}_{gp\times m} \\ \mathbf{O}_{m\times gp} & \left(\frac{1}{\lambda_1\lambda_4}\right) \otimes \mathbf{I}_m \\ \mathbf{O}_{g\times gp} & \mathbf{O}_{g\times m} \end{bmatrix}, \tag{79}$$

where ρ denotes the value of the autoregressive coefficient on first lags in the Minnesota prior, and $\sigma_1, ..., \sigma_n$ denotes, as usual, the standard deviation of the OLS residual obtained from individual autoregressive models. $\mathbf{J}_p$ is defined as

$$\mathbf{J}_p = \text{diag}\left(1^{\lambda_3}, 2^{\lambda_3}, ..., p^{\lambda_3}\right).$$

$\mathbf{Y}_d$ is of dimension $(g(p+1)+m) \times g$, and $\mathbf{X}_d$ is of dimension $(g(p+1)+m) \times (gp+m)$. Considering that each row of $\mathbf{Y}_d$ (or $\mathbf{X}_d$) corresponds to an artificial period, one obtains a total of $T_d = g(p+1)+m$ simulated time periods. Note that unlike the canonical VAR model, $\mathbf{X}_d$ does not correspond to lagged values of $\mathbf{Y}_d$.

Both matrices $\mathbf{Y}_d$ and $\mathbf{X}_d$ are made of three blocks. The first block, made of the first gp rows, is related to the moment of the VAR coefficients corresponding to the endogenous variables of the model. The second block, made of the next m rows, represents the moments of the coefficients on the exogenous variables. Finally, the last block, made of the last g rows, deals with the residual variance covariance matrix.

To make this more concrete, consider a simple example: a VAR model with two endogenous variables and two lags, along with one exogenous variable ($g = 2, m = 1, p = 2$):

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} x_{1,t} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}.$$

For T_d periods, reformulated in the usual stacked form, one obtains:

$$\begin{bmatrix} \frac{\rho\sigma_1}{\lambda_1} & 0 \\ 0 & \frac{\rho\sigma_2}{\lambda_1} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} = \begin{bmatrix} 1^{\lambda_3}\frac{\sigma_1}{\lambda_1} & 0 & 0 & 0 & 0 \\ 0 & 1^{\lambda_3}\frac{\sigma_2}{\lambda_1} & 0 & 0 & 0 \\ 0 & 0 & 2^{\lambda_3}\frac{\sigma_1}{\lambda_1} & 0 & 0 \\ 0 & 0 & 0 & 2^{\lambda_3}\frac{\sigma_2}{\lambda_1} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\lambda_1\lambda_4} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11}^{(1)} & a_{21}^{(1)} \\ a_{12}^{(1)} & a_{22}^{(1)} \\ a_{11}^{(2)} & a_{21}^{(2)} \\ a_{12}^{(2)} & a_{22}^{(2)} \\ c_{11} & c_{21} \end{bmatrix} + \begin{bmatrix} e_{1,1} & e_{2,1} \\ e_{1,2} & e_{2,2} \\ e_{1,3} & e_{2,3} \\ e_{1,4} & e_{2,4} \\ e_{1,5} & e_{2,5} \\ e_{1,6} & e_{2,6} \\ e_{1,7} & e_{2,7} \end{bmatrix}.$$

Note the first block is 2×2 rows, the second is 1 row, and the third is 2 rows.

Let's take a look at the third block (the last two rows), and consider the entries related to the first variable (the first column):

$$\sigma_1 = e_{1,6},$$
$$0 = e_{1,7}.$$

Taking expectations of $e_{1,7}$, one gets

$$\begin{aligned}\mathbb{E}[e_1] &= 0,\\ \text{Var}(e_1) &= \sigma_1^2.\end{aligned}$$

This simply replicates the prior variance for e_1 in the natural conjugate NIW prior.

Now, look at block 1. The first row gives:

$$\begin{aligned}\frac{\rho\sigma_1}{\lambda_1} &= 1^{\lambda_3}\frac{\sigma_1}{\lambda_1}a_{11}^{(1)} + e_{1,1}\\ \implies a_{11}^{(1)} &= \frac{\rho}{1^{\lambda_3}} - \frac{\lambda_1}{1^{\lambda_3}\sigma_1}e_{1,1}\\ \implies \mathbb{E}\left[a_{11}^{(1)}\right] &= \rho,\\ \text{Var}\left(a_{11}^{(1)}\right) &= \lambda_1^2.\end{aligned}$$

The second row gives

$$\begin{aligned}0 &= 1^{\lambda_3}\frac{\sigma_1}{\lambda_1}a_{21}^{(1)} + e_{2,1}\\ \implies a_{21}^{(1)} &= -\frac{\lambda_1}{1^{\lambda_3}\sigma_1}e_{2,1}\\ \implies \mathbb{E}\left[a_{21}^{(1)}\right] &= 0,\\ \text{Var}\left(a_{11}^{(1)}\right) &= \frac{\sigma_2^2}{\sigma_1^2}\lambda_1^2.\end{aligned}$$

Then, look at block 2. Develop the first entry of row 5:

$$\begin{aligned}0 &= \frac{c_{11}}{\lambda_1\lambda_4} + e_{1,5}\\ \implies c_{11} &= -\lambda_1\lambda_4 e_{1,5}\\ \implies \mathbb{E}\left[c_{11}\right] &= 0,\\ \text{Var}\left(c_{11}\right) &= (\lambda_1\lambda_4)^2\sigma_1^2.\end{aligned}$$

Going on the same way with the other entries of blocks 1 and 2, it is straightforward to see that one will recover the full diagonal of the prior variance covariance matrix for $\boldsymbol{\beta}$ implemented in the natural conjugate NIW prior!

But, theres more:

$$\begin{aligned}
\text{Cov}\left(a_{11}^{(1)}, c_{11}\right) &= \mathbb{E}\left[\left(a_{11}^{(1)} - \mathbb{E}[a_{11}^{(1)}]\right)(c_{11} - \mathbb{E}[c_{11}])\right] \\
&= \mathbb{E}\left[a_{11}^{(1)} c_{11} - a_{11}^{(1)}\mathbb{E}[c_{11}] - \mathbb{E}[a_{11}^{(1)}]c_{11} + \mathbb{E}[a_{11}^{(1)}]\mathbb{E}[c_{11}]\right] \\
&= \mathbb{E}\left[a_{11}^{(1)} c_{11} - \rho c_{11}\right] \\
&= \mathbb{E}\left[\left(\frac{\rho}{1^{\lambda_3}} - \frac{\lambda_1}{1^{\lambda_3}\sigma_1} e_{1,1}\right)(-\lambda_1\lambda_4 e_{1,5}) - \rho\left(-\lambda_1\lambda_4 e_{1,5}\right)\right] \\
&= \mathbb{E}\left[-\frac{\rho\lambda_1\lambda_4 e_{1,5}}{1^{\lambda_3}} + \frac{\lambda_1 e_{1,1}\lambda_1\lambda_4 e_{1,5}}{1^{\lambda_3}\sigma_1} + \rho\lambda_1\lambda_4 e_{1,5}\right] \\
&= \lambda_1^2\lambda_4\sigma_1.
\end{aligned}$$

This shows that unlike the normal conjugate NIW prior, the dummy observation setting allows us to implement some prior covariance between the VAR coefficients of the same equation. In this respect, the dummy observation scheme is even richer than the NIW prior.

To conclude our discussion on the basic dummy observation strategy, we show how it combines with the simplified prior introduced at the beginning of the subsection. This is done in a simple way:

$$\begin{aligned}
\mathbf{Y}^* &= \begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}^d \end{bmatrix}, \\
\mathbf{X}^* &= \begin{bmatrix} \mathbf{X} \\ \mathbf{X}^d \end{bmatrix},
\end{aligned}$$

and where $T^* = T + T^d$. That is, $\mathbf{Y}^*$ and $\mathbf{X}^*$ are obtained by concatenating the dummy observation matrices at the top of the actual data matrices, $\mathbf{Y}$ and $\mathbf{X}$. We can then conduct estimation and inference as usual.

Uses for the dummy specification prior include large models when variables are introduced in levels. Because such variables typically include unit roots, the model itself should be characterised by one (or more) unit roots. However, with large models, each draw from the posterior distribution produces VAR coefficients for a large number of equations. This significantly increases the risk that for any draw, at least one equation will obtain coefficients that are actually explosive (have a root greater than one in absolute value) rather than comprising a strict unit root. This may result, for instance, in explosive confidence bands for the IRFs. It would thus be desirable to set a prior that would force the VAR process towards a unit root rather than explosive roots. This can be done in one of two ways:

1. Sum of coefficients approach: As stated above, this forces the VAR process towards a unit root.
2. Initial dummy observation approach: Forces the process towards co-integration. One sets a prior for the model's unconditional mean for the dependent variable. The model remains at

its unconditional mean os that it is stationary despite its unit roots: it must therefore be co-integrated.

4.7 Block exogeneity prior

This concept is closely related to that of Granger causality in standard VAR models. To clarify things, consider again the simple example developed in Section 4.3 when we looked at the mean and variance of the Minnesota prior, with $g = 2$, $p = 2$, and $m = 1$:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} x_{1,t} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}. \tag{80}$$

Suppose that one believes that the second variable does not affect the first variable, that is, it has no impact on it. In terms of the example model, this is as if $y_{1,t}$ is exogenous to $y_{2,t}$, and translates to:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{11}^{(1)} & 0 \\ a_{21}^{(1)} & a_{22}^{(1)} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{11}^{(2)} & 0 \\ a_{21}^{(2)} & a_{22}^{(2)} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} x_{1,t} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}. \tag{81}$$

If the above model is the correct representation of the relation between y_1 and y_2, then we would like to obtain this representation from the posterior of the VAR model. This turns out to be quite easy to implement.

We can set a 0 prior mean on the relevant coefficients, and an arbitrarily small prior variance on them, that way the posterior values will be close to 0 as well. In practice, this means that we would set the prior mean in line with a conventional Minnesota prior, and in vector form $\boldsymbol{\beta}_0$ would be:

$$\boldsymbol{\beta}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

This guarantees that the prior mean of any block exogenous coefficient is 0. Then, we can use the following variance covariance scheme: multiply the block exogenous variance by an additional parameter, λ_5^2, which will be set to an arbitrary small value. This will result in a very tight prior variance on these coefficients. In practice, one may, for example, use the value: $\lambda_5 = 0.001$. Using this strategy

on the above example, one gets:

$$\mathbf{\Omega}_0 = \begin{bmatrix} \lambda_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\sigma_1^2}{\sigma_2^2}(\lambda_1\lambda_2\lambda_5)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\sigma_1^2}{\sigma_2^2}\left(\frac{\lambda_1\lambda_2\lambda_5}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_1^2(\lambda_1\lambda_4)^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\sigma_2^2}{\sigma_1^2}(\lambda_1\lambda_2)^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\sigma_2^2}{\sigma_1^2}\left(\frac{\lambda_1\lambda_2}{2^{\lambda_3}}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_2^2(\lambda_1\lambda_4)^2 \end{bmatrix}$$

Because the prior variance will be very close (in fact it an be made close to 0 by reducing the value of λ_5), the posterior distribution will be extremely tight around 0, as desired.

Of course, block exogeneity need not be limited to one variable only. One may create as many exogenous blocks as required. One need only multiply the prior variance of all the relevant coefficients by λ_5^2 to obtain the desired exogeneity on the posterior mean.

Finally, it should be mentioned that block exogeneity is available with the Minnesota, independent NIW, and normal diffuse priors, but not with the natural conjugate NIW prior nor the dummy observation prior. For the dummy observation prior, the reason is obvious: the prior is diffuse, so $\mathbf{\Sigma} \otimes \mathbf{\Phi}_0$ is simply not defined. For the natural conjugate NIW prior, it is the particular Kronecker structure $\mathbf{\Sigma} \otimes \mathbf{\Phi}_0$ in place of the covariance matrix $\mathbf{\Omega}_0$ that causes instability. This structure implies that the variance of one equation has to be proportional with the variance of the other equations. Hence, imposing block exogeneity on one variable for one equation would lead to imposing it on all the other equations. Not only would it lead to assuming block exogeneity on some equations where it's not desired, but it would also lead to some of the model variables to be exogenous to themselves, which is impossible.

4.8 Time varying parameters VAR

A time-varying parameter VAR (TVP-VAR) model differs from fixed-coefficient VAR models in that they allow the parameters of the model to vary over time, according to a specified law of motion.

The basic TVP-VAR is of the form

$$\mathbf{y}_t = \mathbf{A}_{1,t}\mathbf{y}_{t-1} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{C}_t + \mathbf{u}_t, \tag{82}$$

where the constant coefficients are now replaced by the time-varying $\mathbf{A}_{j,t}$. We can rewrite the above

in compact form as:

$$\mathbf{y}_t = \mathbf{x}_t \boldsymbol{\beta}_t + \mathbf{u}_t, \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\mathbf{x}_t$ is defined similar to (42),

$$\mathbf{x}_t = \mathbf{I}_g \otimes \left[1, \mathbf{y}_{t-1}^\top, ..., \mathbf{y}_{t-p}^\top\right],$$

and

$$\boldsymbol{\beta}_t = \text{vec}\left(\begin{bmatrix} \mathbf{A}_{1,t}^\top \\ \vdots \\ \mathbf{A}_{p,t}^\top \\ \mathbf{C}_t^\top \end{bmatrix}\right).$$

It is common to assume that the coefficients follow a random-walk process:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\varsigma}_t, \tag{83}$$

with

$$\boldsymbol{\varsigma}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Upsilon}),$$

and the initial conditions $\boldsymbol{\beta}^0$ are treated as parameters. Here we make the simplifying assumption that the covariance matrix $\boldsymbol{\Upsilon}$ is restricted to be a diagonal,[11] and the innovations $\boldsymbol{\varsigma}_t$ are uncorrelated with $\mathbf{u}_t$.

The law of motion for $\boldsymbol{\beta}$, (83) – i.e., the state equation – implies that

$$\boldsymbol{\beta}_{t+1} | \boldsymbol{\beta}_t, \boldsymbol{\Upsilon} \sim \mathcal{N}(\boldsymbol{\beta}_t, \boldsymbol{\Upsilon}),$$

which can be used as a prior distribution for $\boldsymbol{\beta}_{t+1}$. Hence, the prior for all the states (i.e., $\boldsymbol{\beta}_t \forall t$) is a product of normal distributions. To complete the model specification, consider independent priors for $\boldsymbol{\Sigma}, \boldsymbol{\beta}^0$, and the diagonal elements of $\boldsymbol{\Upsilon}$;

$$\begin{aligned} \boldsymbol{\Sigma} &\sim \mathcal{W}^{-1}\left(\boldsymbol{\Psi}_0, \nu_0\right), \\ \boldsymbol{\beta}^0 &\sim \mathcal{N}\left(\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0\right), \\ \upsilon_i &\sim \Gamma^{-1}\left(\nu_0^{\upsilon_i}, \Psi_0^{\upsilon_i}\right). \end{aligned}$$

[11] So we have

$$\boldsymbol{\Upsilon} = \text{diag}(\upsilon_1, ..., \upsilon_{g(gp+m)}).$$

4.8.1 TVP-VAR estimation

We now outline an Gibbs sampler to estimate the TVP-VAR model. The model parameters are $\boldsymbol{\beta}^0$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Upsilon}$, and the states are $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, ..., \boldsymbol{\beta}_T^\top)^\top$, We therefore consider a 4-block Gibbs sampler.

First, to sample $\boldsymbol{\beta}$, we rewrite the observation equation, (82), as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_T \otimes \boldsymbol{\Sigma}),$$

and

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & & & \mathbf{O} \\ & \mathbf{x}_2 & & \\ & & \ddots & \\ \mathbf{O} & & & \mathbf{x}_T \end{bmatrix}.$$

Hence, we have

$$\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\mathbf{x}\boldsymbol{\beta}, \mathbf{I}_T \otimes \boldsymbol{\Sigma}\right).$$

So we have reframed the TVP-VAR as a normal linear regression model. Next, we derive the prior for $\boldsymbol{\beta}$. Rewrite the law of motion (83) in matrix notation:

$$\mathbf{H}\boldsymbol{\beta} = \tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}} + \boldsymbol{\varsigma},$$

where

$$\begin{aligned} \boldsymbol{\varsigma} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_T \otimes \boldsymbol{\Upsilon}\right), \\ \tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}} &= \left(\boldsymbol{\beta}_0^\top, \mathbf{0}, ..., \mathbf{0}\right)^\top, \\ \mathbf{H} &= \begin{bmatrix} \mathbf{I}_{g(gp+m)} & & & & & \mathbf{O} \\ -\mathbf{I}_{g(gp+m)} & \mathbf{I}_{g(gp+m)} & & & & \\ & -\mathbf{I}_{g(gp+m)} & \mathbf{I}_{g(gp+m)} & & & \\ & & \ddots & \ddots & & \\ \mathbf{O} & & & & -\mathbf{I}_{g(gp+m)} & \mathbf{I}_{g(gp+m)} \end{bmatrix}. \end{aligned}$$

Note that $\mathbf{H}$ is of dimension $Tg(gp+m) \times Tg(gp+m)$, and is a multivariate generalisation of a first difference matrix. We assume that $|\mathbf{H}| = 1$, and is therefore invertible. We won't cover it here, but one can show that

$$\mathbf{H}^{-1}\tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}} = \mathbf{1}_T \otimes \boldsymbol{\beta}^0.$$

Therefore, the prior of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}|\boldsymbol{\beta}^0, \boldsymbol{\Upsilon} \sim \mathcal{N}\left(\mathbf{1}_T \otimes \boldsymbol{\beta}^0, \left[\mathbf{H}^\top (\mathbf{I}_T \otimes \boldsymbol{\Upsilon}^{-1})\mathbf{H}\right]^{-1}\right).$$

Finally, by standard linear regression results, we obtain

$$\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\Sigma}, \boldsymbol{\beta}^0, \boldsymbol{\Upsilon} \sim \mathcal{N}\left(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\Omega}}\right), \tag{84}$$

where

$$\begin{aligned}
\bar{\boldsymbol{\beta}} &= \bar{\boldsymbol{\Omega}}\left[\mathbf{H}^\top (\mathbf{I}_T \otimes \boldsymbol{\Upsilon}^{-1})\mathbf{H}(\mathbf{1}_T \otimes \boldsymbol{\beta}^0) + \mathbf{X}^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{y}\right], \\
\bar{\boldsymbol{\Omega}} &= \left[\mathbf{H}^\top (\mathbf{I}_T \otimes \boldsymbol{\Upsilon})\mathbf{H} + \mathbf{X}^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{X}\right]^{-1}.
\end{aligned}$$

Next we can show that the marginal distribution for Σ is given by

$$\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Upsilon} \sim \mathcal{W}^{-1}\left(\Psi_0 + \sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta}_t)(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta}_t)^\top, \nu_0 + T\right). \tag{85}$$

In addition, each of the diagonal elements of $\boldsymbol{\Upsilon}$ has an Inverse-Gamma distribution:

$$\upsilon_i|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\beta}^0 \sim \Gamma^{-1}\left(\nu_0^{\upsilon_i} + \frac{T}{2}, \Psi_0^{\upsilon_i} + \frac{1}{2}\sum_{t=1}^{T}(\beta_{it} - \beta_{i(t-1)})^2\right). \tag{86}$$

Finally, since $\boldsymbol{\beta}^0$ only appears in the first state equation:

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}^0 + \boldsymbol{\varsigma}_1, \quad \boldsymbol{\varsigma}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Upsilon}).$$

Given the normal prior, $\boldsymbol{\beta}^0 \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0)$, we can use standard linear regression results to get

$$\boldsymbol{\beta}^0|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Upsilon} \sim \mathcal{N}\left(\bar{\boldsymbol{\beta}}^0, \bar{\boldsymbol{\Omega}}^0\right),$$

with

$$\begin{aligned}
\bar{\boldsymbol{\beta}}^0 &= \bar{\boldsymbol{\Omega}}^0\left(\boldsymbol{\Omega}_0\boldsymbol{\beta}_0 + \boldsymbol{\Upsilon}^{-1}\boldsymbol{\beta}_1\right), \\
\bar{\boldsymbol{\Omega}}^0 &= \boldsymbol{\Omega}_0 + \boldsymbol{\Upsilon}.
\end{aligned}$$

The Gibbs sampler is summarised as follows: Pick some initial values for $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$, $\boldsymbol{\Upsilon}^{(0)}$, and $\boldsymbol{\beta}^{0(0)}$, then repeat the following steps from $s = 1, ..., S$:

1. Draw $\boldsymbol{\beta}^{(s)} \sim \left(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\Sigma}^{(s-1)}, \boldsymbol{\Upsilon}^{(s-1)}, \boldsymbol{\beta}^{0(s-1)}\right)$ (multivariate normal)

2. Draw $\boldsymbol{\Sigma}^{(s)} \sim \left(\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta}^{(s)}, \boldsymbol{\Upsilon}^{(s-1)}, \boldsymbol{\beta}^{0(s-1)}\right)$ (Inverse-Wishart)
3. Draw $\boldsymbol{\Upsilon}^{(s)} \sim \left(\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta}^{(s)}, \boldsymbol{\Sigma}^{(s)}, \boldsymbol{\beta}^{0(s-1)}\right)$ (independent Inverse-Gammas)
4. Draw $\boldsymbol{\beta}^{0(s)} \sim \left(\boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\beta}^{(s)}, \boldsymbol{\Sigma}^{(s)}, \boldsymbol{\Upsilon}^{(s)}\right)$ (multivariate normal)

4.9 VAR with stochastic volatility

When stochastic volatility is added to the framework (referred to as an SV-VAR), the VAR innovations are assumed to still be normally distributed, but with variance that evolves over time. We start with a constant coefficient VAR and write it as a linear regression in which the errors have a time varying covariance matrix, $\boldsymbol{\Sigma}_t$:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{u}_t, \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t). \tag{87}$$

Ideally, we want $\boldsymbol{\Sigma}_t$ to evolve smoothly, while at each time period $\boldsymbol{\Sigma}_t$ is a valid variance covariance matrix – i.e., it is positive definite and symmetric. This methods follows that of Cogley and Sargent (2005).

The idea is to model $\boldsymbol{\Sigma}_t$ as

$$\boldsymbol{\Sigma}_t^{-1} = \mathbf{P}^\top \boldsymbol{\Lambda}_t^{-1} \mathbf{P},$$

where $\boldsymbol{\Lambda}_t$ is a diagonal matrix and $\mathbf{P}$ is a lower triangular matrix with ones on the main diagonal:

$$\boldsymbol{\Lambda}_t = \begin{bmatrix} \exp(h_{1,t}) & 0 & \cdots & 0 \\ 0 & \exp(h_{2,t}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \exp(h_{g,t}) \end{bmatrix},$$

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ p_{21} & 1 & 0 & \cdots & 0 \\ p_{31} & p_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{g1} & p_{g2} & \cdots & p_{g(g-1)} & 1 \end{bmatrix}.$$

By construction, $\boldsymbol{\Sigma}_t$ is symmetric and positive definite for any values of $\mathbf{h}_t = (h_{1,t}, ..., _{g,t})^\top$ and $\mathbf{p} = (p_{21}, p_{31}, p_{32}, ..., p_{g1}, ..., p_{g(g-1)})^\top$. Note that the dimension of $\mathbf{p}$ is $g(g-1)/2$. Then, each $h_{i,t}$ is specified independently using a univariate stochastic volatility model. More precisely, each $h_{i,t}$ evolves according to the following random walk:

$$h_{i,t} = h_{i(t-1)} + u_{i,t}^h, \quad u_{i,t} \sim \mathcal{N}(0, \sigma_{h,i}^2),$$

and $h_{i,0}$ is treated as an unknown parameter.

In contrast, the parameters $\mathbf{p}$ are restricted to be constant here.[12]

By construction $\mathbf{\Sigma}_t = \mathbf{P}^{-1}\mathbf{\Lambda}_t(\mathbf{P}^{-1})^\top$, and therefore we can express each element of $\mathbf{\Sigma}_t$ in terms of the elements of $\mathbf{\Lambda}_t$ and

$$\mathbf{P}^{-1} = (p^{ij}).$$

More precisely, we have

$$\sigma_{ii,t} = \exp(h_{i,t}) + \sum_{k=1}^{i-1} \exp(h_{k,t})(p^{ik})^2,\ i = 1, ..., n = gp + m.,$$
$$\sigma_{ij,t} = p^{ij}\exp(h_{j,t}) + \sum_{k=1}^{j-1} p^{ik}p^{jk}\exp(h_{k,t}),\ 1 \leq j < i \leq n,$$

where $\sigma_{ij,t}$ is the (i,j) element of $\mathbf{\Sigma}_t$. In particular, the log-volatility $h_{1,t}$ affects the variances of all the variables, whereas $h_{n,t}$ impacts only the last variable.

In addition, despite the assumption of a constant matrix $\mathbf{P}$, this setup allows for some form of time-varying correlations among the innovations. This can be seen via a simple example. Using the formulas above, we have

$$\sigma_{11,t} = \exp(h_{1,t}),$$
$$\sigma_{22,t} = \exp(h_{2,t}) + \exp(h_{1,t})(p^{21})^2,$$
$$\sigma_{12,t} = p^{21}\exp(h_{1,t}).$$

We have used the fact that $\mathbf{P}^{-1}$ is a lower triangular matrix with ones on the main diagonal, and therefore $p^{11} = 1$ and $p^{12} = 0$. Now, the $(1,2)$ correlation coefficient is given by

$$\frac{\sigma_{12,t}}{(\sigma_{11,t}\sigma_{22,t})^{1/2}} = \frac{p^{12}}{(\exp(h_{2,t} - h_{1,t}) + (p^{21})^2)^{1/2}}.$$

Hence, as long as $h_{1,t}$ and $h_{2,t}$ are not identical for all t, this correlation coefficient is time varying.

To complete the model specification, consider independent priors for $\boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\sigma}_h^2 = (\sigma_{h,1}^2, ..., \sigma_{h,n}^2)^\top$,

[12] Primiceri (2005) considers an extension where these parameters are time varying and modelled as random walks. It turns out all that is needed is an extra block to sample these time varying parameters from a linear Gaussian state space model.

and $\mathbf{h}_0 = (h_{10}, ..., h_{n0})^\top$:

$$\begin{aligned}
\boldsymbol{\beta} &\sim \mathcal{N}\left(\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0\right), \\
\mathbf{p} &\sim \mathcal{N}\left(\mathbf{p}_0, \boldsymbol{\Omega}_0^{\mathbf{p}}\right), \\
\sigma_{h,i}^2 &\sim \Gamma^{-1}\left(\nu_0^{h_i}, \Psi_0^{h_i}\right), \\
\mathbf{h}_0 &\sim \mathcal{N}\left(\mathbf{b}_0, \mathbf{B}_0\right).
\end{aligned}$$

4.9.1 Stochastic volatility VAR estimation

To estimate the SV-VAR model, we describe a Gibbs sampler below. The model parameters $\boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\sigma}_h^2, \mathbf{h}_0$, and the states are the log-volatility $\mathbf{h}_{i,1:T} = (h_{i1}, ..., h_{iT})^\top$. Hence, we consider a 5-block Gibbs sampler. The two key steps are sampling $\mathbf{a}$ and $\mathbf{h} = (\mathbf{h}_{1,1:T}^\top, ..., \mathbf{h}_{n,1:T}^\top)^\top$.

Begin with the sample of $\mathbf{p}$, the lower triangular elements of $\mathbf{P}$. First observe that given $\mathbf{y}$ and $\boldsymbol{\beta}$, $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is known. Then, we rewrite the model as a system of regressions in which u_{it} is regressed on the negative values of $u_{1,t}, ..., u_{(i-1),t}$ for $i = 2, ..., n$ and $p_{i1}, ..., p_{i(i-1)}$ are the corresponding regression coefficients. If we can rewrite the model this way, then we can apply standard linear regression results to sample $\mathbf{p}$.

Note:

$$\begin{aligned}
\mathbf{P}\mathbf{u}_t &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ p_{21} & 1 & 0 & \cdots & 0 \\ p_{31} & p_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{g1} & p_{g2} & \cdots & p_{g(g-1)} & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \\ \vdots \\ u_{n,t} \end{bmatrix} = \begin{bmatrix} u_{1,t} \\ u_{2,t} + p_{21}u_{1,t} \\ u_{3,t} + p_{31}u_{1,t} + p_{32}u_{2,t} \\ \vdots \\ u_{n,t} + \sum_{j=1}^{n-1} p_{nj}u_{j,t} \end{bmatrix} \\
&= \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \\ \vdots \\ u_{n,t} \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ -u_{1,t} & 0 & 0 & 0 & 0 & \cdots & \cdots & \vdots \\ 0 & -u_{1,t} & -u_{2,t} & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & & & \ddots & \ddots & & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 & -u_{1,t} & \cdots & -u_{t,(n-1)} \end{bmatrix} \begin{bmatrix} p_{21} \\ p_{31} \\ p_{32} \\ \vdots \\ p_{n(n-1)} \end{bmatrix}
\end{aligned}$$

$\Leftrightarrow \mathbf{P}\mathbf{u}_t = \boldsymbol{u}_t - \boldsymbol{\Theta}_t\mathbf{p}$.

Noting that $|\boldsymbol{\Sigma}_t| = |\boldsymbol{\Lambda}_t|$, we can write the likelihood implied by (87) as

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{p}, \mathbf{h}) &\propto \left(\prod_{t=1}^{T} |\boldsymbol{\Lambda}_t|^{-1/2}\right) \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} \mathbf{u}_t^\top (\mathbf{P}^\top \boldsymbol{\Lambda}_t^{-1} \mathbf{P}) \mathbf{u}_t\right\} \\ &= \left(\prod_{t=1}^{T} |\boldsymbol{\Lambda}_t|^{-1/2}\right) \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} (\mathbf{P}\mathbf{u}_t)^\top \boldsymbol{\Lambda}_t^{-1} \mathbf{P}\mathbf{u}_t\right\} \\ &= \left(\prod_{t=1}^{T} |\boldsymbol{\Lambda}_t|^{-1/2}\right) \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} (\boldsymbol{u}_t - \boldsymbol{\Theta}_t \mathbf{p})^\top \boldsymbol{\Lambda}_t^{-1} (\boldsymbol{u}_t - \boldsymbol{\Theta}_t \mathbf{p})\right\}. \end{aligned} \tag{88}$$

In other words, the likelihood is the same as that implied by the regression

$$\mathbf{u}_t = \boldsymbol{\Theta}_t \mathbf{p} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_t).$$

Therefore stacking these over $t = 1, ..., T$ we get:

$$\mathbf{u} = \boldsymbol{\Theta}\mathbf{p} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}),$$

with

$$\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}_1, ..., \boldsymbol{\Lambda}_T).$$

Given the prior $\mathbf{p} \sim \mathcal{N}(\mathbf{p}_0, \boldsymbol{\Omega}_0^{\mathbf{p}})$, it then follows that

$$\mathbf{p}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{h} \sim \mathcal{N}(\bar{\mathbf{p}}, \bar{\boldsymbol{\Omega}}^{\mathbf{p}}), \tag{89}$$

where

$$\begin{aligned} \bar{\mathbf{p}} &= \bar{\boldsymbol{\Omega}}^{\mathbf{p}} \left(\boldsymbol{\Omega}_0^{\mathbf{p}} \mathbf{p}_0 + \boldsymbol{\Theta}^\top \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{u}\right), \\ \bar{\boldsymbol{\Omega}}^{\mathbf{p}} &= \boldsymbol{\Omega}_0^{\mathbf{p}} + \boldsymbol{\Theta}^{-1} \boldsymbol{\Lambda} (\boldsymbol{\Theta}^{-1})^\top. \end{aligned}$$

To sample the log-volatility, $\mathbf{h}$, we first compute the orthogonalised innovations:

$$\begin{aligned} \tilde{\mathbf{u}}_t &= \mathbf{P}(\mathbf{y}_t - \mathbf{X}_t \boldsymbol{\beta}), \ t = 1, ..., T, \\ \implies \mathbb{E}[\tilde{\mathbf{u}}_t | \mathbf{p}, \mathbf{h}, \boldsymbol{\beta}] &= \mathbf{0}, \\ \text{Var}(\tilde{\mathbf{u}}_t | \mathbf{p}, \mathbf{h}, \boldsymbol{\beta}) &= \mathbf{P}(\mathbf{P}^\top \boldsymbol{\Lambda}_t^{-1} \mathbf{P})^{-1} \mathbf{P}^\top = \boldsymbol{\Lambda}_t, \\ \therefore \tilde{u}_{i,t} | \mathbf{p}, \mathbf{h}, \boldsymbol{\beta} &\sim \mathcal{N}(0, \exp\{h_{i,t}\}). \end{aligned} \tag{90}$$

The other steps are now standard. For example, to sample $\boldsymbol{\beta}$, we rewrite (87) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}),$$

and $\tilde{\boldsymbol{\Sigma}} = \text{diag}\left(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_T\right)$ is a block-diagonal matrix. Together with the prior, $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0)$, we have

$$\boldsymbol{\beta}|\mathbf{y}, \mathbf{p}, \mathbf{h} \sim \mathcal{N}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\Omega}}), \tag{91}$$

where

$$\begin{aligned}
\bar{\boldsymbol{\beta}} &= \bar{\boldsymbol{\Omega}}\left(\boldsymbol{\Omega}_0\boldsymbol{\beta}_0 + \mathbf{X}^\top\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{y}\right), \\
\bar{\boldsymbol{\Omega}} &= \boldsymbol{\Omega}_0 + \mathbf{X}^{-1}\tilde{\boldsymbol{\Sigma}}(\mathbf{X}^{-1})^\top,
\end{aligned}$$

and remember that

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}} &= \text{diag}\left(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_T\right), \\
\boldsymbol{\Sigma}_t &= \mathbf{P}^{-1}\boldsymbol{\Lambda}_t(\mathbf{P}^{-1})^{-1}.
\end{aligned}$$

4.10 Bayesian panel VARs

A panel VAR describes the the evolution of $\mathbf{y}_{t,i}$, the vector of $g \times 1$ endogenous variables of each unit $i \in [1, ..., N]$, by a system of p-th order VARs. In its most general form, the panel VAR model for unit i is written as:

$$\mathbf{y}_{t,i} = \sum_{j=1}^{N}\sum_{k=1}^{p}\mathbf{A}_{ij,t}^{k}\mathbf{y}_{j,t-k} + \mathbf{C}_{ij}\mathbf{x}_t + \mathbf{u}_{i,t},$$

with

$$\begin{aligned}
\mathbf{u}_{i,t} &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{ii,t}), \\
\mathbf{\Sigma}_{ii,t} &= \mathbb{E}\left[\mathbf{u}_{i,t}\mathbf{u}_{i,t}^{\top}\right] \\
&= \mathbb{E}\left[\begin{bmatrix} u_{i,1,t} \\ u_{i,2,t} \\ \vdots \\ u_{i,g,t} \end{bmatrix} \begin{bmatrix} u_{i,1,t} & u_{i,2,t} & \cdots & u_{i,g,t} \end{bmatrix}\right] \\
&= \begin{bmatrix} \sigma_{i,1,t}^2 & \sigma_{i,2,t}\sigma_{i,2,t} & \cdots & \sigma_{i,1,t}\sigma_{i,g,t} \\ \sigma_{i,2,t}\sigma_{i,1,t} & \sigma_{i,2,t}^2 & \cdots & \sigma_{i,2,t}\sigma_{i,g,t} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{i,g,t}\sigma_{i,1,t} & \sigma_{i,g,t}\sigma_{i,2,t} & \cdots & \sigma_{i,g,t}^2 \end{bmatrix}.
\end{aligned}$$

$\mathbf{u}_{i,t}$ is assumed to be non-autocorrelated, so that $\mathbb{E}\left[\mathbf{u}_{i,t}\mathbf{u}_{i,t}^{\top}\right] = \mathbf{\Sigma}_{ii,t}$, while $\mathbb{E}\left[\mathbf{u}_{i,t}\mathbf{u}_{i,s}^{\top}\right] = \mathbf{O}$ when $t \neq s$. Note that in this general setting the variance covariance matrix for the VAR residuals is allowed be period specific, which implies a general form of heteroskedasticity.

For each variable in unit i, the dynamic equation at period t contains a total of $k = Ngp + m$ coefficients to estimate, implying $q = g(Ngp + m)$ coefficients to estimate for the whole unit. Stacking over the N units, the model can be reformulated as

$$\begin{aligned}
\mathbf{y}_t &= \sum_{k=1}^{p} \mathbf{A}_t^k \mathbf{y}_{t-k} + \mathbf{C}_t \mathbf{x}_t + \mathbf{u}_t, \\
&= \mathbf{A}_t^1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_t^p \mathbf{y}_{t-p} + \mathbf{C}_t \mathbf{x}_t + \mathbf{u}_t,
\end{aligned} \tag{92}$$

or

$$\underbrace{\begin{bmatrix}\mathbf{y}_{1,t}\\ \mathbf{y}_{2,t}\\ \vdots\\ \mathbf{y}_{N,t}\end{bmatrix}}_{Ng\times 1} = \underbrace{\begin{bmatrix}\mathbf{A}_{11,t}^{(1)} & \mathbf{A}_{12,t}^{(1)} & \cdots & \mathbf{A}_{1N,t}^{(1)}\\ \mathbf{A}_{21,t}^{(1)} & \mathbf{A}_{22,t}^{(1)} & \cdots & \mathbf{A}_{2N,t}^{(1)}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{A}_{N1,t}^{(1)} & \mathbf{A}_{N2,t}^{(1)} & \cdots & \mathbf{A}_{NN,t}^{(1)}\end{bmatrix}}_{Ng\times Ng}\begin{bmatrix}\mathbf{y}_{1,t-1}\\ \mathbf{y}_{2,t-1}\\ \vdots\\ \mathbf{y}_{N,t-1}\end{bmatrix} + \cdots$$
$$+ \begin{bmatrix}\mathbf{A}_{11,t}^{(p)} & \mathbf{A}_{12,t}^{(p)} & \cdots & \mathbf{A}_{1N,t}^{(p)}\\ \mathbf{A}_{21,t}^{(p)} & \mathbf{A}_{22,t}^{(p)} & \cdots & \mathbf{A}_{2N,t}^{(p)}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{A}_{N1,t}^{(p)} & \mathbf{A}_{N2,t}^{(p)} & \cdots & \mathbf{A}_{NN,t}^{(p)}\end{bmatrix}\begin{bmatrix}\mathbf{y}_{1,t-p}\\ \mathbf{y}_{2,t-p}\\ \vdots\\ \mathbf{y}_{N,t-p}\end{bmatrix} + \underbrace{\begin{bmatrix}\mathbf{C}_{1,t}\\ \mathbf{C}_{2,t}\\ \vdots\\ \mathbf{C}_{N,t}\end{bmatrix}}_{Ng\times m}\mathbf{x}_t + \underbrace{\begin{bmatrix}\mathbf{u}_{1,t}\\ \mathbf{u}_{2,t}\\ \vdots\\ \mathbf{u}_{N,t}\end{bmatrix}}_{Ng\times 1}.$$

The vector of residuals, $\mathbf{u}_t$, has the following properties:

$$\mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t),$$

with

$$\begin{aligned}\mathbf{\Sigma}_t &= \mathbb{E}\left[\mathbf{u}_t\mathbf{u}_t^\top\right]\\ &= \mathbb{E}\left[\begin{bmatrix}\mathbf{u}_{1,t}\\ \mathbf{u}_{2,t}\\ \vdots\\ \mathbf{u}_{N,t}\end{bmatrix}\begin{bmatrix}\mathbf{u}_{1,t} & \mathbf{u}_{2,t} & \cdots & \mathbf{u}_{N,t}\end{bmatrix}\right]\\ &= \underbrace{\begin{bmatrix}\mathbf{\Sigma}_{11,t} & \mathbf{\Sigma}_{12,t} & \cdots & \mathbf{\Sigma}_{1N,t}\\ \mathbf{\Sigma}_{21,t} & \mathbf{\Sigma}_{22,t} & \cdots & \mathbf{\Sigma}_{2N,t}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{\Sigma}_{N1,t} & \mathbf{\Sigma}_{N2,t} & \cdots & \mathbf{\Sigma}_{NN,t}\end{bmatrix}}_{Ng\times Ng}.\end{aligned}$$

The assumption of absence of autocorrelation is then extended to the whole model: $\mathbb{E}\left[\mathbf{u}_t\mathbf{u}_t^\top\right] = \mathbf{\Sigma}_t$ and $\mathbb{E}\left[\mathbf{u}_t\mathbf{u}_s^\top\right] = \mathbf{O},\ t \neq s$. Now there are $h = Nq = Ng(Ngp + m)$ coefficients to estimate.

This is the most general form of the panel VAR model. Under this form, it is characterised by four properties:

1. Dynamic interdependencies: the dynamic behaviour of each unit is determined by lagged values of itself, but also by lagged values of all the other endogenous variables of all other units. In

other words, $\mathbf{A}_{ij,t}^{k} \neq \mathbf{O}$ when $i \neq j$.

2. Static interdependencies: the $\mathbf{u}_{i,t}$ are allowed to be correlated across units. That is, in general, $\mathbf{\Sigma}_{ij,t} \neq \mathbf{O}$ when $i \neq j$.

3. Cross-subsectional heterogeneity: the VAR coefficients and residual variances are allowed to be unit-specific.

4. Dynamic heterogeneity: the VAR coefficients and the residual variance covariance matrix are allowed to be period specific. In other words, $\mathbf{A}_{ij,t}^{k} \neq \mathbf{A}_{ij,s}^{k}$ and $\mathbf{\Sigma}_{ij,t} \neq \mathbf{\Sigma}_{ij,s}$ when $t \neq s$.

In practice, this general form may be too complex to yield accurate estimates, as it consumes many degrees of freedom. If one has legitimate reasons to assume that some of the properties will not hold, better estimates can be obtained by relaxing them and opt for less degrees of freedom consuming procedures.

4.10.1 Bayesian Panel VAR examples

Suppose we had $N = 2$ (US and EU), $g = 2$ variables (GDP and interest rates), over $p = 1$ lag, and $m = 1$ exogenous variable (some constant), then we could write:

$$\begin{bmatrix} y_t^{US} \\ r_t^{US} \\ y_t^{EU} \\ r_t^{EU} \end{bmatrix} = \begin{bmatrix} \beta_{11,t} & \beta_{12,t} & \beta_{13,t} & \beta_{14,t} \\ \beta_{21,t} & \beta_{22,t} & \beta_{23,t} & \beta_{24,t} \\ \beta_{31,t} & \beta_{32,t} & \beta_{33,t} & \beta_{34,t} \\ \beta_{41,t} & \beta_{42,t} & \beta_{43,t} & \beta_{44,t} \end{bmatrix} \begin{bmatrix} y_{t-1}^{US} \\ r_{t-1}^{US} \\ y_{t-1}^{EU} \\ r_{t-1}^{EU} \end{bmatrix} + \begin{bmatrix} c_{11,t} \\ c_{21,t} \\ c_{31,t} \\ c_{41,t} \end{bmatrix} x_{1,t} + \begin{bmatrix} u_{1,t}^{US} \\ u_{2,t}^{US} \\ u_{1,t}^{EU} \\ u_{2,t}^{EU} \end{bmatrix},$$

with

$$\begin{aligned} \mathbf{\Sigma}_t &= \begin{bmatrix} \mathbf{\Sigma}_{US,EU,t} & \mathbf{\Sigma}_{US,US,t} \\ \mathbf{\Sigma}_{EU,US,t} & \mathbf{\Sigma}_{EU,EU,t} \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} & \sigma_{13,t} & \sigma_{14,t} \\ \sigma_{21,t} & \sigma_{22,t} & \sigma_{23,t} & \sigma_{24,t} \\ \sigma_{31,t} & \sigma_{32,t} & \sigma_{33,t} & \sigma_{34,t} \\ \sigma_{41,t} & \sigma_{42,t} & \sigma_{43,t} & \sigma_{44,t} \end{bmatrix}, \end{aligned}$$

and the four properties are:

1. Dynamic interdependencies (red terms): Variables are allowed to be determined by lagged values of other variables, e.g., $\beta_{11,t} \neq 0$.

2. Static interdependencies (green terms): Residuals are allowed to be correlated across countries, e.g., $\sigma_{13,t} \neq 0$.

3. Cross-subsectional heterogeneity (blue terms): the VAR coefficients and residual variances are allowed to be country specific, e.g., $\beta_{11,t} \neq \beta_{33,t}$.

4. Dynamic heterogeneity: the VAR coefficients and the residual variance covariance matrix are allowed to be time varying. In other words, $\mathbf{A}^k_{ij,t} \neq \mathbf{A}^k_{ij,s}$ and $\mathbf{\Sigma}_{ij,t} \neq \mathbf{\Sigma}_{ij,s}$ when $t \neq s$.

As mentioned, integrating all of these 4 properties is often not optimal. We need to reduce the number of properties, or to find a way to simplify the problem.

Case 1: No property satisfied

$$\begin{bmatrix} y_t^{US} \\ r_t^{US} \\ y_t^{EU} \\ r_t^{EU} \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & 0 & 0 \\ \beta_{21} & \beta_{22} & 0 & 0 \\ 0 & 0 & \beta_{11} & \beta_{12} \\ 0 & 0 & \beta_{21} & \beta_{22} \end{bmatrix} \begin{bmatrix} y_{t-1}^{US} \\ r_{t-1}^{US} \\ y_{t-1}^{EU} \\ r_{t-1}^{EU} \end{bmatrix} + \begin{bmatrix} c_{11} \\ c_{21} \\ c_{31} \\ c_{41} \end{bmatrix} x_{1,t} + \begin{bmatrix} u_t^{US} \\ u_t^{US} \\ u_t^{EU} \\ u_t^{EU} \end{bmatrix}, \tag{93}$$

with

$$\mathbf{\Sigma}_t = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 & 0 \\ \sigma_{21} & \sigma_{22} & 0 & 0 \\ 0 & 0 & \sigma_{11} & \sigma_{12} \\ 0 & 0 & \sigma_{21} & \sigma_{22} \end{bmatrix}. \tag{94}$$

Note that the coefficients and variance covariance terms are not longer time varying, the coefficient and variance covariance matrices are now a diagonal block matrix, and that the error terms within each country is the same across variables.

Case 2: Cross-section heterogeneity

$$\begin{bmatrix} y_t^{US} \\ r_t^{US} \\ y_t^{EU} \\ r_t^{EU} \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & 0 & 0 \\ \beta_{21} & \beta_{22} & 0 & 0 \\ 0 & 0 & \beta_{33} & \beta_{34} \\ 0 & 0 & \beta_{43} & \beta_{44} \end{bmatrix} \begin{bmatrix} y_{t-1}^{US} \\ r_{t-1}^{US} \\ y_{t-1}^{EU} \\ r_{t-1}^{EU} \end{bmatrix} + \begin{bmatrix} c_{11} \\ c_{21} \\ c_{31} \\ c_{41} \end{bmatrix} x_{1,t} + \begin{bmatrix} u_t^{US} \\ u_t^{US} \\ u_t^{EU} \\ u_t^{EU} \end{bmatrix}, \tag{95}$$

with

$$\mathbf{\Sigma}_t = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 & 0 \\ \sigma_{21} & \sigma_{22} & 0 & 0 \\ 0 & 0 & \sigma_{33} & \sigma_{34} \\ 0 & 0 & \sigma_{43} & \sigma_{44} \end{bmatrix}. \tag{96}$$

Notice that the the block diagonal matrices now feature different elements.

Case 3: Dynamic and static interdependencies

$$\begin{bmatrix} y_t^{US} \\ r_t^{US} \\ y_t^{EU} \\ r_t^{EU} \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \\ \beta_{31} & \beta_{32} & \beta_{33} & \beta_{34} \\ \beta_{41} & \beta_{42} & \beta_{43} & \beta_{44} \end{bmatrix} \begin{bmatrix} y_{t-1}^{US} \\ r_{t-1}^{US} \\ y_{t-1}^{EU} \\ r_{t-1}^{EU} \end{bmatrix} + \begin{bmatrix} c_{11} \\ c_{21} \\ c_{31} \\ c_{41} \end{bmatrix} x_{1,t} + \begin{bmatrix} u_t^{US} \\ u_t^{US} \\ u_t^{EU} \\ u_t^{EU} \end{bmatrix}, \tag{97}$$

with

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}. \tag{98}$$

As the name suggests, we now have cross-period interdependencies.

Case 4: Dynamic heterogeneity

$$\begin{bmatrix} y_t^{US} \\ r_t^{US} \\ y_t^{EU} \\ r_t^{EU} \end{bmatrix} = \begin{bmatrix} \beta_{11,t} & \beta_{12,t} & \beta_{13,t} & \beta_{14,t} \\ \beta_{21,t} & \beta_{22,t} & \beta_{23,t} & \beta_{24,t} \\ \beta_{31,t} & \beta_{32,t} & \beta_{33,t} & \beta_{34,t} \\ \beta_{41,t} & \beta_{42,t} & \beta_{43,t} & \beta_{44,t} \end{bmatrix} \begin{bmatrix} y_{t-1}^{US} \\ r_{t-1}^{US} \\ y_{t-1}^{EU} \\ r_{t-1}^{EU} \end{bmatrix} + \begin{bmatrix} c_{11,t} \\ c_{21,t} \\ c_{31,t} \\ c_{41,t} \end{bmatrix} x_{1,t} + \begin{bmatrix} u_t^{US} \\ u_t^{US} \\ u_t^{EU} \\ u_t^{EU} \end{bmatrix}, \tag{99}$$

with

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} & \sigma_{13,t} & \sigma_{14,t} \\ \sigma_{21,t} & \sigma_{22,t} & \sigma_{23,t} & \sigma_{24,t} \\ \sigma_{31,t} & \sigma_{32,t} & \sigma_{33,t} & \sigma_{34,t} \\ \sigma_{41,t} & \sigma_{42,t} & \sigma_{43,t} & \sigma_{44,t} \end{bmatrix}. \tag{100}$$

References

An, S. and Schorfheide, F. (2007), "Bayesian Analysis of DSGE Models", *Econometric Reviews*, 26/2-4: 113–72.

Christiano, L. J. and Eichenbaum, M. S. (1992), "Current Real-Business-Cycle Theories and Aggregate Labor-Market Fluctuations", *American Economic Review*, 82/3: 430–50.

Christiano, L. J., Eichenbaum, M. S., and Evans, C. L. (2005), "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy", *Journal of Political Economy*, 113/1: 1–45.

Christiano, L. J., Trabandt, M., and Walentin, K. (2011), "Introducing Financial Frictions and Unemployment into a Small Open Economy Model", *Journal of Economic Dynamics and Control*, 35/12: 1999–2041.

Cogley, T. and Sargent, T. J. (2005), "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII US", *Review of Economic Dynamics*, 8/2: 262–302.

Davidson, R. and MacKinnon, J. G. (2004), *Econometric Theory and Methods* (Oxford University Press).

DeJong, D. N. and Dave, C. (2012), *Structural Macroeconometrics* (2nd Edition, Princeton University Press).

Enders, W. (2010), *Applied Econometric Time Series* (3rd Edition, John Wiley and Sons).

Fernández-Villaverde, J. (2010), "The Econometrics of DSGE Models", *SERIEs*: 3–49.

Hansen, L. P. (1982), "Large Sample Properties of Generalised Method of Moments Estimators", *Econometrica*, 50/4: 1029–54.

Ireland, P. N. (2004), "A Method for Taking Models to the Data", *Journal of Economic Dynamics and Control*, 28/6: 1205–26.

Kim, J. (2000), "Constructing and Estimating a Realistic Optimizing Model of Monetary Policy", *Journal of Monetary Economics*, 45/2: 329–59.

Kydland, F. E. and Prescott, E. C. (1982), "Time to Build and Aggregate Fluctuations", *Econometrica*, 50/6: 1345–70.

Litterman, R. B. (1980), *Bayesian Procedure for Forecasting with Vector Autoregressions* (MIT Press).

Litterman, R. B. (1986), "Forecasting with Bayesian Vector Autoregressions – Five Years of Experience", *Journal of Business and Economic Statistics*, 4/1: 25–38.

Miao, J. (2020), *Economic Dynamics in Discrete Time* (2nd Edition, MIT Press).

Primiceri, G. E. (2005), "Time Varying Structural Vector Autoregressions and Monetary Policy", *The Review of Economic Studies*, 72/3: 821–52.

Rotemberg, J. J. and Woodford, M. (1997), "An Optimisation-Based Econometric Framework for the Evaluation of Monetary Policy", *NBER Macroeconomics Annual*, 12: 297–361.

Sims, C. A. (1980), "Macroeconomics and Reality", *Econometrica*, 48/1: 1–48.

Smets, F. and Wouters, R. (2003), "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area", *Journal of the European Economic Association*, 1/5: 1123–75.

Smets, F. and Wouters, R. (2007), "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach", *American Economic Review*, 97/3: 586–606.

Taylor, J. B. (1993), "Discretion Verses Policy Rules in Practice", *Carnegie-Rochester Conference Series on Public Policy*: 195–214.