– Lecture Notes 1

An introduction to Bayesian Vector Autoregressions

These lecture notes contain derivations used in the BVAR set of slides and a discussion of certain aspects of Bayesian estimation. We first define some useful probability distributions that will be used throughout the notes. After these initial definitions, we begin with the Bayesian analysis of the linear regression model. Then, we discuss more general Bayesian VARs.

These notes are not a complete treatment of Bayesian econometrics. If you are interested in this topic, you should consult some of the many books on Bayesian econometrics.

Preliminaries

In this section, we define the multivariate normal density, the Wishart density, and the inverse Wishart density. We also discuss Bayes's Theorem, which is the most important tool in Bayesian analysis. Throughout this note, we denote the determinant of a matrix A by $|A|$ and the trace of a square matrix B by $\mathrm{tr}(B)$.[1] We also use $p(\cdot)$ to denote the probability of an event or the probability density function of a random variable or vector, depending on the context.

To warm up, we start by defining the univariate normal and the Gamma distributions. The Gamma distribution will also be used when we consider the Bayesian analysis of the simple linear regression model.

Definition 1.1 — Normal random variable

Let X be a continuous random variable defined for all real numbers. We say that X has a **normal distribution** with mean $\mu \in R$ and variance $\sigma^2 \in R_+$, denoted by $X \sim N(\mu, \sigma^2)$, if its probability density function is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}. \tag{1.1}$$

When $\mu = 0$ and $\sigma^2 = 1$ we obtain the standard normal distribution.

With normal random variables, we can construct other random variables of interest. For example, if we consider the sum of squares of n independent standard normal variables, we

[1] Recall that the trace of a square matrix is defined as the sum of the diagonal elements.

obtain the chi-square distribution, which is often used in the estimation of variance and hypothesis testing. In particular, if Y_i for $i = 1, \dots, n$ are independent standard normal random variables (so that $Y_i \sim N(0, 1)$), the random variable $X = Y_1^2 + Y_2^2 + ...Y_n^2$ has a chi-square distribution with n degrees of freedom, often written as $X \sim \chi^2_n$. The chi-square distribution is a special case of the Gamma distribution.

The Gamma distribution is widely used in Bayesian analysis as a prior distribution for parameters that must be positive, such as variances. The Gamma distribution depends on two parameters, a shape parameter k and a rate parameter θ. Above, we mentioned that the sum of squares of independent standard normal random variables is distributed as a chi-square. For the case of arbitrary variances, the sum of squares now follows a Gamma distribution. In particular, suppose that $Y_i \sim N(0, \sigma^2)$ for $i = 1, 2, \dots, n$ are independent normal random variables. Then, $X = Y_1^2 + Y_2^2 + ...Y_n^2$ is distributed as a Gamma random variable with shape parameter $k = n/2$ and rate parameter $\theta = 1/(2\sigma^2)$.

The Gamma distribution is defined as follows:

Definition 1.2 — Gamma distribution

Let X be a continuous random variable defined for $X \geq 0$. We say that X has a **Gamma** distribution with shape parameter $k > 0$ and rate parameter $\theta > 0$, denoted by $X \sim \mathit{Gamma}(k, \theta)$, if its probability density function is

$$f(x|k,\theta) = \frac{\theta^k}{\Gamma(k)} x^{k-1} e^{-\theta x}, \quad x \geq 0, \tag{1.2}$$

where $\Gamma(k) = \int_0^\infty t^{k-1} e^{-t}\, dt$ is the Gamma function.

We now consider multivariate generalizations of the previous formulas. We start with the multivariate normal distribution and next define the Wishart distribution, which is a generalization of the Gamma distribution to the multivariate case. It will also become useful below to define the inverse Wishart distribution, which is the distribution of the inverse of a Wishart.

Definition 1.3 — Multivariate Normal distribution

Let $\boldsymbol{X} \in R^k$ be a random vector with k elements, $\boldsymbol{X} = (X_1, X_2, \dots, X_k)'$, and let $\boldsymbol{\mu} \in R^k$, and Σ be a positive definite symmetric $k \times k$ matrix. We say that $\boldsymbol{X}$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix Σ, denoted by $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if the probability density function of $\boldsymbol{X}$ is given by

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \tag{1.3}$$

We now consider the multivariate generalization of the sum of squares of mean-zero normal random variables. Now we have to consider all possible cross products between the random variables.

Definition 1.4 — Wishart distribution

Let $X_i \in R^l$ for $i = 1, 2, \dots, m$ be an $l-$dimensional i.i.d. random vector distributed according to $X_i \sim N(\mathbf{0}_{l\times 1}, \Sigma)$, where Σ is an $l \times l$ positive definite matrix. Let $Z = \sum_{i=1}^{m} X_i X_i'$, which is an $l \times l$ random matrix with the sum of all squared terms in X_i. The distribution of Z is called a ($l-$dimensional) **Wishart distribution** with scale matrix Σ and m degrees of freedom. The probability density function of Z is given by

$$f(Z|\Sigma, m) = \frac{|Z|^{(m-l-1)/2}}{2^{ml/2}\,|\Sigma|^{m/2}\,\Gamma_l(m/2)} \exp\left[-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}Z)\right] \tag{1.4}$$

where $\Gamma_l(t) = \pi^{l(l-1)/4}\prod_{i=1}^{l}\Gamma(t-(i-1)/2)$ is the multivariate Gamma function. We write $Z \sim W_l(\Sigma, m)$

If the random matrix Z follows a Wishart distribution, Z^{-1} follows an Inverse Wishart distribution:

Definition 1.5 — Inverse Wishart distribution

We say that Z follows an **inverse Wishart distribution** with scale matrix Ψ and m degrees of freedom, denoted by $Z \sim iW(\Psi, m)$, if Z^{-1} has a Wishart distribution with scale matrix Ψ^{-1} and m degrees of freedom, so that $Z^{-1} \sim W(\Psi^{-1}, m)$. The probability density function of the inverse Wishart is given by

$$f(Z|\Psi, m) = \frac{|\Psi|^{m/2}\,|Z|^{(m+l+1)/2}}{2^{ml/2}\,|\Sigma|^{m/2}\,\Gamma_l(m/2)} \exp\left[-\frac{1}{2}\mathrm{tr}(\Psi Z^{-1})\right]. \tag{1.5}$$

For completeness, we also state the Bayes Theorem:

Theorem 1.1 — Bayes' rule

Let x and y be random variables (or vectors). Then

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}. \tag{1.6}$$

Proof. The proof follows from the definition of conditional probabilities:

$$\begin{aligned} p(x|y) =& \frac{p(x,y)}{p(y)} \Rightarrow p(x,y) = p(x|y)p(y), \\ p(y|x) =& \frac{p(x,y)}{p(x)} \Rightarrow p(x,y) = p(y|x)p(x). \end{aligned}$$

Equating both terms and rearranging gives the first equality in equation (1.6). To obtain the second equality use the Law of total probability: $\int p(x|y)p(y)dy = \int p(x,y)dy = p(x)$. ■

Bayesian analysis of the linear regression model

Consider the standard linear regression model

$$y_i = x_i'\beta + \epsilon_i \tag{1.7}$$

where $i = 1, 2, ..., n$ is the number of observations, y_i is the dependent variable, x_i is a vector with k regressors, $\beta \in R^k$ and $\epsilon_i \sim N(0, \sigma^2)$.

Stacking the observations in (1.7) we obtain

$$Y = X\beta + \epsilon; \qquad \epsilon \sim N(0, \sigma^2 I_n), \tag{1.8}$$

where Y is $n \times 1$, X is $n \times k$ and ϵ is $n \times 1$.

Since $\epsilon = Y - X\beta$ is distributed as a multivariate normal, the likelihood function of (1.8) is

$$\begin{aligned} p(Y|\beta, \sigma^2) =& (2\pi)^{-\frac{n}{2}} \left|\sigma^2 I_n\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(Y - X\beta)' \left(\sigma^2 I_n\right)^{-1} (Y - X\beta)\right) \\ =& (2\pi)^{-\frac{n}{2}} \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right). \end{aligned} \tag{1.9}$$

Consider first the Maximum Likelihood Estimator (MLE) of β. To that end, take logs on both sides of (1.9) and obtain the log-likelihood function

$$\begin{aligned} l(Y|\beta, \sigma^2) =& -\frac{n}{2}\log(2\pi) + \log\left((\sigma^2)^{-\frac{n}{2}}\right) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \\ =& -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left[Y'Y - 2\beta'X'Y + \beta'X'X\beta\right]. \end{aligned}$$

Take the first order condition with respect to β and rearrange to obtain

$$\hat{\beta} = (X'X)^{-1} X'Y.$$

That is, the MLE estimator is identical to the OLS estimator of β.

We can write

$$\hat{\beta} = (X'X)^{-1} X'Y = (X'X)^{-1} X'(X\beta + \epsilon) = \beta + (X'X)^{-1} X'\epsilon.$$

and note that

$$E\left[\hat{\beta}\right] = \beta$$

and

$$\begin{aligned} E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\right] =& E\left[(X'X)^{-1} X'\epsilon\epsilon'X\left((X'X)^{-1}\right)'\right] \\ =& (X'X)^{-1} X'E\left[\epsilon\epsilon'\right] X\left((X'X)^{-1}\right)' \\ =& (X'X)^{-1} X'\sigma^2 I_T X\left((X'X)^{-1}\right)' \\ =& \sigma^2 (X'X)^{-1} (X'X)\left((X'X)^{-1}\right)' \\ =& \sigma^2 \left((X'X)'\right)^{-1} \\ =& \sigma^2 (X'X)^{-1}. \end{aligned}$$

It then follows that

$$\hat{\beta} \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right).$$

The MLE relies solely on data for parameter estimation. On the other hand, Bayesian analysis offers the ability to incorporate prior beliefs about the parameters through prior density distributions for β and σ^2. In our case, we aim to perform a Bayesian analysis where we treat β and σ^2 as random variables. Specifically, our objective is to determine the posterior distribution of β and σ^2 after observing the data Y. To achieve this, we combine prior distributions of β and σ^2 with the information contained in the data using the likelihood function $p(Y|\beta, \sigma^2)$. The result is the posterior distribution $p(\beta, \sigma^2|Y)$.

However, instead of directly obtaining the joint posterior $p(\beta, \sigma|Y)$, we will compute the conditional posterior probabilities $p(\beta|\sigma^2, Y)$ and $p(\sigma^2|\beta, Y)$. In this example, we introduce the **Gibbs Sampler**, a Markov Chain Monte Carlo sampling technique used to draw samples from the joint posterior $p(\beta, \sigma|Y)$. It relies on the knowledge of the conditional posteriors $p(\beta|\sigma^2, Y)$ and $p(\sigma^2|\beta, Y)$, which we will derive. First, we consider the case where σ^2 is known, and derive a closed form expression of the conditional posterior $p(\beta|\sigma^2, y)$. Then, assuming knowledge of β, we derive the closed form expression of the conditional posterior $p(\sigma^2|\beta, Y)$. Finally, we explain how to use the Gibbs Sampler to draw samples from the joint posterior $p(\beta, \sigma^2|Y)$.

Bayesian analysis with unknown β and known σ^2

Let's pretend for the moment that we know the variance σ^2 of the error term ϵ_i. This is equivalent to saying that we are conditioning on a given value of σ^2.

We start with a prior belief about β in the form of a distribution $p(\beta)$, which we assume to be normal:

$$\beta \sim N(\beta_0, \Sigma_0),$$

where β_0 is $k \times 1$ and Σ_0 is $k \times k$ —there are k regressors in the linear model (1.7).

Next, we construct the likelihood function $p(Y|\beta, \sigma^2)$ as above. Finally, we combine the prior belief with the likelihood function to obtain the posterior distribution of the parameters $p(\beta|\sigma^2, Y)$ given the data.

Bayes's theorem then implies

$$\begin{aligned} p(\beta|Y, \sigma^2) =& \frac{p(Y|\beta, \sigma^2)p(\beta)}{\int_\beta p(Y|\beta, \sigma^2)p(\beta)d\beta} \\ \propto& p(Y|\beta, \sigma^2)p(\beta), \end{aligned}$$

where the symbol $\propto$ means "is proportional to".

The prior density of the parameter β is given by

$$\begin{aligned} p(\beta) =& (2\pi)^{-\frac{k}{2}} |\Sigma_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\beta - \beta_0\right)' \Sigma_0^{-1} \left(\beta - \beta_0\right)\right) \\ \propto& \exp\left[-\frac{1}{2}\left(\beta - \beta_0\right)' \Sigma_0^{-1} \left(\beta - \beta_0\right)\right], \end{aligned}$$

Since we assume that σ^2 is known, we can write the likelihood function (1.9) as proportional to

$$p\left(Y|\beta,\sigma^2\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left(Y-X\beta\right)'\left(Y-X\beta\right)\right)$$

because $(\sigma^2)^{n/2}$ is a constant that we assume known and, therefore, we can ignore.

Combining the prior with the likelihood function, we obtain the kernel of the posterior distribution:

$$\begin{aligned} p(\beta|Y,\sigma^2) \propto & p(Y|\beta,\sigma^2)p(\beta) \\ & \propto \exp\left[-\frac{1}{2\sigma^2}\left(Y-X\beta\right)'\left(Y-X\beta\right)\right]\exp\left[-\frac{1}{2}\left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)\right] \\ & \propto \exp\left[-\frac{1}{2}\left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)-\frac{1}{2\sigma^2}\left(Y-X\beta\right)'\left(Y-X\beta\right)\right] \\ & \propto \exp\left[-\frac{1}{2}\left(\beta-\beta_1\right)'\Sigma_1^{-1}\left(\beta-\beta_1\right)\right], \end{aligned}$$

where

$$\begin{aligned} \Sigma_1^{-1} = & \left(\Sigma_0^{-1}+\frac{1}{\sigma^2}X'X\right) \\ \beta_1 = & \Sigma_1\left(\Sigma_0^{-1}\beta_0+\frac{1}{\sigma^2}X'Y\right). \end{aligned}$$

Now, since the previous equation is the kernel of a normal distribution, we know that the conditional posterior is normally distributed, so that

$$\beta|\{Y,\sigma^2\} \sim N\left(\beta_1,\Sigma_1\right). \tag{1.10}$$

That is, the conditional posterior $p(\beta|Y,\sigma^2)$ is (multivariate) normally distributed with mean β_1 and variance Σ_1.

We now prove the previous claim. Consider the term inside the exponential function

$$Z = -\frac{1}{2}\left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)-\frac{1}{2\sigma^2}\left(Y-X\beta\right)'\left(Y-X\beta\right).$$

We want to write this expression in the form

$$Z = -\frac{1}{2}\left(\beta-\beta_1\right)'\Sigma_1^{-1}\left(\beta-\beta_1\right)+\text{a constant},$$

for some β_1 and Σ_1 that are yet to be found. This trick is called "completing the square" and is used a lot. First note that

$$\left(\beta-\beta_1\right)'\Sigma_1^{-1}\left(\beta-\beta_1\right)=\beta'\Sigma_1^{-1}\beta-2\beta_1'\Sigma_1^{-1}\beta+\beta_1'\Sigma_1^{-1}\beta_1. \tag{1.11}$$

Now we work with the term inside the exponent of the posterior kernel

$$\begin{aligned} & \left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)+\frac{1}{\sigma^2}\left(Y-X\beta\right)'\left(Y-X\beta\right) \\ & =\beta'\Sigma_0^{-1}\beta-2\beta_0'\Sigma_0^{-1}\beta+\beta_0'\Sigma_0^{-1}\beta_0+\frac{1}{\sigma^2}Y'Y-\frac{2}{\sigma^2}Y'X\beta+\frac{1}{\sigma^2}\beta'X'X\beta \\ & =\beta'\left(\Sigma_0^{-1}+\frac{1}{\sigma^2}X'X\right)\beta-2\left(\beta_0'\Sigma_0^{-1}+\frac{2}{\sigma^2}Y'X\right)\beta+\beta_0'\Sigma_0^{-1}\beta_0+\frac{1}{\sigma^2}Y'Y. \end{aligned} \tag{1.12}$$

Compare this last expression with the right side of equation (1.11). They are quite similar. Let's define

$$\Sigma_1^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2} X'X$$

and

$$\beta_1' \Sigma_1^{-1} = \beta_0' \Sigma_0^{-1} + \frac{2}{\sigma^2} Y'X \Rightarrow \beta_1 = \Sigma_1 \left(\Sigma_0^{-1} \beta_0 + \frac{2}{\sigma^2} X'Y \right).$$

With these definitions, equation (1.12) can be written as

$$\begin{aligned} &\beta' \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} X'X \right) \beta - 2 \left(\beta_0' \Sigma_0^{-1} + \frac{2}{\sigma^2} Y'X \right) \beta + \beta_0' \Sigma_0^{-1} \beta_0 + \frac{1}{\sigma^2} Y'Y \\ =&\beta' \Sigma_1^{-1} \beta - 2\beta_1' \Sigma_1^{-1} \beta + \beta_1' \Sigma_1^{-1} \beta_1 - \beta_1' \Sigma_1^{-1} \beta_1 + \beta_0' \Sigma_0^{-1} \beta_0 + \frac{1}{\sigma^2} Y'Y \\ =& \left(\beta - \beta_1 \right)' \Sigma_1^{-1} \left(\beta - \beta_1 \right) + \beta_0' \Sigma_0^{-1} \beta_0 - \beta_1' \Sigma_1^{-1} \beta_1 + \frac{1}{\sigma^2} Y'Y \\ =& \left(\beta - \beta_1 \right)' \Sigma_1^{-1} \left(\beta - \beta_1 \right) + C \end{aligned}$$

where C is a constant (i.e. does not depend on the random variable β). This term is like (1.11) for the proposed β_1 and Σ_1 with a constant in the exponent. Also, note that in the second line we added and subtracted the term $\beta_1' \Sigma_1^{-1} \beta_1$.

Summarizing, we showed that the kernel of the posterior is

$$p(\beta | Y, \sigma^2) \propto \exp \left[-\frac{1}{2} \left(\beta - \beta_1 \right)' \Sigma_1^{-1} \left(\beta - \beta_1 \right) \right]$$

which proves that the posterior is normal $\beta | \left\{ Y, \sigma^2 \right\} \sim N \left(\beta_1, \Sigma_1 \right)$ with

$$\begin{aligned} \Sigma_1^{-1} =& \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} X'X \right) \\ \beta_1 =& \Sigma_1 \left(\Sigma_0^{-1} \beta_0 + \frac{1}{\sigma^2} X'Y \right). \end{aligned}$$

But also note that the OLS estimator is $\hat{\beta} = \left(X'X \right)^{-1} X'Y \Rightarrow X'Y = \left(X'X \right) \hat{\beta}$. Therefore,

$$\begin{aligned} \beta_1 =& \Sigma_1 \left(\Sigma_0^{-1} \beta_0 + \frac{1}{\sigma^2} \left(X'X \right) \hat{\beta} \right) \\ =& \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} X'X \right)^{-1} \left(\Sigma_0^{-1} \beta_0 + \frac{1}{\sigma^2} \left(X'X \right) \hat{\beta} \right) \\ =& W \times \beta_0 + \left(I - W \right) \times \hat{\beta}, \end{aligned}$$

where $W = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} X'X \right)^{-1} \Sigma_0^{-1}$.

That is, the mean of the posterior distribution of β is a weighted average of the prior mean, β_0, and of the OLS estimator $\hat{\beta}$. Note that if the variance of the prior goes to infinity (e.g. $\Sigma_0 = m\tilde{\Sigma}_0$ with $m \to \infty$, then $\Sigma_0^{-1} \to 0$) we get that $W \to \mathbf{0}$ and the mean of the posterior is the OLS estimate $\beta_1 = \hat{\beta}$. On the other hand, if the data is uninformative, so that $\sigma \to \infty$, we get $W \to I$, $\beta_1 \to \beta_0$ and the mean of the posterior is the mean of the prior.

In addition, note that setting $\Sigma_0^{-1} = \frac{\lambda}{\sigma^2} I$ and $\beta_0 = 0$ gives the Ridge regression:[2]

$$\begin{aligned}\beta_1 &= \left(\lambda I + X'X\right)^{-1}\left(X'X\right)\hat{\beta} \\ &= \left(\lambda I + X'X\right)^{-1} X'Y.\end{aligned}$$

Example 1: Suppose that the linear regression is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0,1).$$

Simulate data assuming $\beta_0 = -1$, $\beta_1 = 2$, $n = 3$ and $x_i \sim \text{Uniform}(0,1)$. Note that the variance of ϵ_i is fixed at 1. Consider two priors for β:

a) Uniform prior $p(\beta) \propto 1$.

b) Normal prior $p(\beta) \sim N(0, I_2)$.

See the slides for the posteriors in this example.

Exercise: Derive the posterior for the uniform prior for β.

Bayesian analysis with unknown σ^2 and known β

We now reverse the roles and assume that β is known and the variance parameter σ^2 is unknown. We aim to estimate the conditional posterior of σ^2. A typical prior for the (inverse of the) variance of a linear regression is the Gamma distribution (1.2) that we discussed in the Preliminaries section.

Imposing a prior to the variance σ^2 is equivalent to imposing a prior to the precision parameter $\tau \equiv 1/\sigma^2$. We will assume that prior of the precision is a Gamma distribution $\tau \sim Gamma(a_0, b_0)$. The prior distribution of τ is thus proportional to

$$p(\tau) \propto \tau^{a_0 - 1} \exp\left(-\tau b_0\right).$$

The mean of τ is $E(\tau) = a_0/b_0$ and the variance of τ is $\text{Var}(\tau) = a_0/b_0^2$. The hyperparameters a_0 and b_0 determine the shape of the Gamma distribution.

Using τ instead of σ^2, we can rewrite the kernel of the likelihood function (1.9) as

$$p(Y|\beta,\tau) \propto \tau^{\frac{n}{2}} \exp\left[-\frac{\tau}{2}\left(Y - X\beta\right)'\left(Y - X\beta\right)\right]$$

If we combine the kernel of the prior with the kernel of the likelihood, we obtain the kernel of the posterior distribution for τ:

$$\begin{aligned}p(\tau|\beta,Y) \propto& p(Y|\tau,\beta)p(\tau) \\ \propto& \tau^{\frac{n}{2}}\tau^{a_0-1}\exp\left[-b_0\tau\right]\exp\left[-\frac{\tau}{2}\left(Y - X\beta\right)'\left(Y - X\beta\right)\right] \\ \propto& \tau^{a_0+\frac{n}{2}-1}\exp\left[-\tau\left(b_0 + \frac{1}{2}\left(Y - X\beta\right)'\left(Y - X\beta\right)\right)\right]\end{aligned}$$

[2] If you don't know what a Ridge regression is, don't worry, you can ignore this paragraph.

Note that this is the kernel of a Gamma distribution

$$\tau|\{\beta, Y\} \sim Gamma(a_1, b_1) \tag{1.13}$$

with parameters

$$\begin{aligned} a_1 =& a_0 + \frac{n}{2}, \\ b_1 =& b_0 + \frac{1}{2}(Y - X\beta)'(Y - X\beta). \end{aligned}$$

So if we know β and the prior for the precision is a Gamma distribution, we know that the conditional posterior of the precision is also a Gamma distribution but with different hyperparameters.

Bayesian analysis for unknown β and σ^2

Given a normal prior for β and a Gamma prior for $\tau = 1/\sigma^2$, we were able to compute the conditional posteriors $p(\beta|\tau, Y)$ and $p(\tau|\beta, Y)$. But we want to sample from the joint posterior distribution $p(\beta, \tau|Y)$. How do we do it?

In this case there is a powerful tool to draw samples from $p(\beta, \tau|Y)$ given that we know the two conditional posteriors $p(\beta|\tau, Y)$ and $p(\tau|\beta, Y)$. This tool is the Gibbs Sampler. Specifically, we will construct a Markov chain $\{\beta^j, \tau^j\}$ for $j = 1, 2, \ldots, N$ for large N that will converge to the target posterior density $p(\beta, \tau|Y)$. We are not going to prove this result, but it can be shown that the Gibbs Sampler generates a Markov chain whose invariant distribution is precisely the posterior $p(\beta, \tau|Y)$. The algorithm is as follows:[3]

Algorithm: Gibbs Sampler for the linear regression model: Choose a large N, an arbitrary τ^0, and set $j = 1$. Then iterate on the following loop:

```
a) Draw $\beta^j$ from $p(\beta|\tau^{j-1})$
b) Draw $\tau^j$ from $p(\tau|\beta^j)$
c) Store $\{\beta^j, \tau^j\}$, set $j = j + 1$, and return to step a) while $j < N$.
```

Bayesian Vector Autoregressions

Consider the vector autoregression (VAR)

$$Y_t = c + D_1 Y_{t-1} + D_2 Y_{t-2} + \ldots + D_p Y_{t-p} + v_t \tag{1.14}$$

where Y_t is $n \times 1$ (n time series), v_t is an $n \times 1$ vector of errors such that $v_t \sim N(0, \Omega)$, c is an $n \times 1$ vector, and D_j are $n \times n$ coefficient matrices. The (unrestricted) $VAR(p)$ is a multiple equation regression in which the regressors of each equation are the p lagged values of all the variables in Y. In particular, there are n equations, each of which has $k = np + 1$ regressors (the 1 is the constant) so there is a total of $kn = n^2p + 1$ coefficients in the constant and

[3] See the slides for the results.

the D_j matrices. We assume we have data for $t = 1, 2, ..., T$ taking the initial observations $y_{-p+1}, y_{-p+2}, ..., y_0$ as given. In other words, we have $T + p$ observations and let T be the number of effective observations for each variable in Y.

Transpose the VAR(p) (1.14) and write it as

$$\begin{aligned} Y_t' =& Y_{t-1}'D_1' + Y_{t-2}'D_2' + \ldots + Y_{t-p}'D_p' + c' + v_t' \\ =& \begin{bmatrix} Y_{t-1}' & Y_{t-2}' & \cdots & Y_{t-p}' & c' \end{bmatrix} \begin{bmatrix} D_1' \\ D_2' \\ \vdots \\ D_p' \\ c' \end{bmatrix} + v_t' \\ =& X_t\boldsymbol{\beta} + v_t', \end{aligned}$$

where $X_t = \begin{bmatrix} Y_{t-1}' & Y_{t-2}' & \cdots & Y_{t-p}' & 1 \end{bmatrix}$ is $1 \times (np + 1)$ and

$$\underbrace{\boldsymbol{\beta}}_{(np+1)\times n} = \begin{bmatrix} D_1' \\ D_2' \\ \vdots \\ D_p' \\ c' \end{bmatrix}.$$

Let $k \equiv np + 1$, so that X_t is $1 \times k$ and $\boldsymbol{\beta}$ is $k \times n$ matrix. Note that

$$Y_t' = X_t\boldsymbol{\beta} + v_t' \tag{1.15}$$

is a system of n equations, one for each time period $t = 1, 2, ..., T$.

Now we stack vertically all the variables across observations. In particular, let

$$\boldsymbol{Y}_{T\times n} = \begin{bmatrix} Y_1' \\ Y_2' \\ \vdots \\ Y_T' \end{bmatrix}; \quad \boldsymbol{X}_{T\times k} = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_T' \end{bmatrix}; \quad \boldsymbol{V}_{T\times n} = \begin{bmatrix} v_1' \\ v_2' \\ \vdots \\ v_T' \end{bmatrix}$$

Then we can write the system of equations simultaneously for all periods as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{V}. \tag{1.16}$$

Now vectorize the equations and note that $\boldsymbol{X\beta} = \boldsymbol{X\beta}I_n \Rightarrow \text{vec}(\boldsymbol{X\beta}) = \text{vec}(\boldsymbol{X\beta}I_n) = (I_n \otimes \boldsymbol{X})\,\text{vec}(\boldsymbol{\beta})$. Then,

$$\text{vec}(\boldsymbol{Y}) = (I_n \otimes \boldsymbol{X})\,\text{vec}(\boldsymbol{\beta}) + \text{vec}(\boldsymbol{V}). \tag{1.17}$$

Let

$$y_{nT\times 1} = \text{vec}(\boldsymbol{Y});\ \beta_{kn\times 1} = \text{vec}(\boldsymbol{\beta});\ \text{and}\ v_{nT\times 1} = \text{vec}(\boldsymbol{V}) \sim N(0, \Omega \otimes I_T).$$

Then we can write the VAR(p) (1.14) in the format of a standard linear regression model for a single dependent variable:

$$y = (I_n \otimes \boldsymbol{X})\,\beta + v. \tag{1.18}$$

Exercise: Prove that $E(vv') = \Omega \otimes I_T$. (This is ugly but straightforward algebra).

We will now write the likelihood function in two different (but equivalent) ways that will prove useful later on. Since $v_{nT\times 1} \sim N(0, \Omega \otimes I_T)$, we know that

$$y \sim N((I_n \otimes \boldsymbol{X})\beta, \Omega \otimes I_T).$$

Hence, the likelihood function is

$$L(\beta, \Omega|Y) = \frac{|\Omega \otimes I_T|^{-1/2}}{(2\pi)^{nT/2}} \exp\left[-\frac{1}{2}(y - (I_n \otimes \boldsymbol{X})\beta)' (\Omega \otimes I_T)^{-1} (y - (I_n \otimes \boldsymbol{X})\beta)\right]. \quad (1.19)$$

Now, use that for arbitrary matrices $A_{n\times n}$ and $B_{m\times m}$, $\det(A \otimes B) = \det(A)^m \det(B)^n$. Then we can write

$$|\Omega \otimes I_T| = |\Omega|^T.$$

Also, the property $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ implies $(\Omega \otimes I_T)^{-1} = \Omega^{-1} \otimes I_T^{-1}$. Using those two results, we write the likelihood as

$$L(\beta, \Omega|Y) = \frac{|\Omega|^{-T/2}}{(2\pi)^{nT/2}} \exp\left[-\frac{1}{2}(y - (I_n \otimes \boldsymbol{X})\beta)' \left(\Omega^{-1} \otimes I_T\right) (y - (I_n \otimes \boldsymbol{X})\beta)\right] \quad (1.20)$$

Now write the likelihood in a different, but equivalent, way that uses the matrix form (1.16). To that end, we use the following results about matrices, vectorization, and kronecker products: for comformable matrices B, C, D, we have

$$\begin{aligned}
\text{vec}(BCD) &= (D' \otimes B)\,\text{vec}(C) && (1.21)\\
\text{tr}(B'C) &= \text{vec}(B')\,\text{vec}(C) && (1.22)\\
\text{tr}(BCD) &= \text{tr}(CDB) = \text{tr}(DBC) && (1.23)
\end{aligned}$$

where $\text{vec}(\cdot)$ is the vectorization operator and $\text{tr}(\cdot)$ is the trace function.

Using that $y - (I_n \otimes \boldsymbol{X})\beta = v = \text{vec}(\boldsymbol{V})$, we can write the quadratic form in the exponent of (1.20) as

$$\begin{aligned}
(y - (I_n \otimes \boldsymbol{X})\beta)' \left(\Omega^{-1} \otimes I_T^{-1}\right) (y - (I_n \otimes \boldsymbol{X})\beta) &= \text{vec}(\boldsymbol{V})' \left(\Omega^{-1} \otimes I_T\right) \text{vec}(\boldsymbol{V})\\
&= \text{vec}(\boldsymbol{V})' \,\text{vec}\left(I_T \boldsymbol{V} \Omega^{-1}\right)\\
&= \text{vec}(\boldsymbol{V})' \,\text{vec}\left(\boldsymbol{V} \Omega^{-1}\right)\\
&= \text{tr}\left(\boldsymbol{V}'\boldsymbol{V}\Omega^{-1}\right)\\
&= \text{tr}\left(\Omega^{-1}\boldsymbol{V}'\boldsymbol{V}\right) && (1.24)
\end{aligned}$$

where in the second equality we used (1.21), the fourth equality uses (1.22), and the last equality uses (1.23). Then we can equivalently write the likelihood function as

$$L(\beta, \Omega|Y) = \frac{|\Omega|^{-T/2}}{(2\pi)^{nT/2}} \exp\left[-\frac{1}{2}\text{tr}\left(\Omega^{-1}\boldsymbol{V}'\boldsymbol{V}\right)\right]. \quad (1.25)$$

MLE estimation of β

Rewrite the quadratic term in (1.20) as follows:

$$\begin{aligned}
&\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)'\left(\Omega^{-1}\otimes I_T\right)\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)\\
&=y'\left(\Omega^{-1}\otimes I_T\right)y-2\beta'\left(I_n\otimes\boldsymbol{X}'\right)\left(\Omega^{-1}\otimes I_T\right)y+\beta'\left(I_n\otimes\boldsymbol{X}'\right)\left(\Omega^{-1}\otimes I_T\right)\left(I_n\otimes\boldsymbol{X}\right)\beta\\
&=y'\left(\Omega^{-1}\otimes I_T\right)y-2\beta'\left(\Omega^{-1}\otimes\boldsymbol{X}'\right)y+\beta'\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)\beta
\end{aligned}$$

where the last equality uses

$$\left(I_n\otimes\boldsymbol{X}'\right)\left(\Omega^{-1}\otimes I_T\right)=\left(\Omega^{-1}\otimes\boldsymbol{X}'\right)$$

and

$$\left(I_n\otimes\boldsymbol{X}'\right)\left(\Omega^{-1}\otimes I_T\right)\left(I_n\otimes\boldsymbol{X}\right)=\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)$$

Taking logs in (1.20), we obtain the log-likelihood function

$$l\left(\beta,\Omega|Y\right)=-\frac{nT}{2}\log\left(2\pi\right)-\frac{T}{2}\log\left|\Omega\right|-\frac{1}{2}\left[y'\left(\Omega^{-1}\otimes I_T\right)y-2\beta'\left(\Omega^{-1}\otimes\boldsymbol{X}'\right)y+\beta'\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)\beta\right].$$

To obtain the MLE estimator of β, take the first order condition of $l\left(\beta,\Omega|Y\right)$ and equate to zero:

$$\left(\Omega^{-1}\otimes\boldsymbol{X}'\right)y-\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)\beta^{\text{mle}}=0. \tag{1.26}$$

Solving for β^{mle},

$$\beta^{\text{mle}}=\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)^{-1}\left(\Omega^{-1}\otimes\boldsymbol{X}'\right)y.$$

But using the property $\left(A\otimes B\right)^{-1}=A^{-1}\otimes B^{-1}$, we have $\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)^{-1}=\Omega\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, so that

$$\beta^{\text{mle}}=\left(\Omega\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(\Omega^{-1}\otimes\boldsymbol{X}'\right)y.$$

Finally, using $\left(A\otimes B\right)\left(C\otimes D\right)=AC\otimes BC$ for comformable matrices, we have $\left(\Omega\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(\Omega^{-1}\otimes\boldsymbol{X}'\right)$ $\left(\Omega\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)=I_n\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right).^{-1}\boldsymbol{X}'$. Hence, the MLE estimator is

$$\beta^{\text{mle}}=\left(I_n\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)y \tag{1.27}$$

Note that the MLE estimator coincides with the OLS estimator of equation (1.18), which minimizes the quadratic equation

$$\begin{aligned}
Q^{\text{ols}}\left(\beta\right)&=\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)'\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)\\
&=y'y-2\beta'\left(I_n\otimes\boldsymbol{X}\right)'y+\beta'\left(I_n\otimes\boldsymbol{X}\right)'\left(I_n\otimes\boldsymbol{X}\right)\beta\\
&=y'y-2\beta'\left(I_n\otimes\boldsymbol{X}'\right)y+\beta'\left(I_n\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)\beta.
\end{aligned}$$

The first order condition is

$$-2\left(I_n\otimes\boldsymbol{X}'\right)y+2\left(I_n\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)\beta^{\text{ols}}=0\Rightarrow\beta^{\text{ols}}=\left(I_n\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)^{-1}\left(I_n\otimes\boldsymbol{X}'\right)y$$

or

$$\beta^{\text{ols}} = \left(I_n \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)y. \tag{1.28}$$

Exercise: the GLS estimator minimizes the quadratic expression

$$Q^{\text{gls}}(\beta) = (y - (I_n \otimes \boldsymbol{X})\beta)'(\Omega \otimes I_T)^{-1}(y - (I_n \otimes \boldsymbol{X})\beta).$$

Prove that $\beta^{\text{gls}} = \beta^{\text{mle}} = \beta^{\text{ols}}$.

Yet another way of writing the likelihood

We now prove yet another way of writing the likelihood function that will be useful below. First note that

$$\Omega^{-1} \otimes I_T = \left(\Omega^{-1/2} \otimes I_T\right)\left(\Omega^{-1/2} \otimes I_T\right)$$

and that

$$\left(\Omega^{-1/2} \otimes I_T\right)' = \Omega^{-1/2} \otimes I_T.$$

We now do some algebra with the quadratic term in (1.20)

$$\begin{aligned}
& (y - (I_n \otimes \boldsymbol{X})\beta)'\left(\Omega^{-1} \otimes I_T\right)(y - (I_n \otimes \boldsymbol{X})\beta) \\
= \; & (y - (I_n \otimes \boldsymbol{X})\beta)'\left(\Omega^{-1/2} \otimes I_T\right)\left(\Omega^{-1/2} \otimes I_T\right)(y - (I_n \otimes \boldsymbol{X})\beta) \\
= \; & \left[\left(\Omega^{-1/2} \otimes I_T\right)'(y - (I_n \otimes \boldsymbol{X})\beta)\right]'\left(\Omega^{-1/2} \otimes I_T\right)(y - (I_n \otimes \boldsymbol{X})\beta) \\
= \; & \left[\left(\Omega^{-1/2} \otimes I_T\right)(y - (I_n \otimes \boldsymbol{X})\beta)\right]'\left(\Omega^{-1/2} \otimes I_T\right)(y - (I_n \otimes \boldsymbol{X})\beta) \\
= \; & \left[\left(\Omega^{-1/2} \otimes I_T\right)y - \left(\Omega^{-1/2} \otimes I_T\right)(I_n \otimes \boldsymbol{X})\beta\right]'\left(\left(\Omega^{-1/2} \otimes I_T\right)y - \left(\Omega^{-1/2} \otimes I_T\right)(I_n \otimes \boldsymbol{X})\beta\right) \\
= \; & \left[\left(\Omega^{-1/2} \otimes I_T\right)y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\beta\right]'\left[\left(\Omega^{-1/2} \otimes I_T\right)y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\beta\right]
\end{aligned}$$

where in the third line we used $(AB)' = B'A'$ for $A = (y - (I_n \otimes \boldsymbol{X})\beta)$ and $B = \left(\Omega^{-1/2} \otimes I_T\right)$ and the other properties discussed above.

Now rewrite the term

$$\begin{aligned}
\left(\Omega^{-1/2} \otimes I_T\right)y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\beta &= \left(\Omega^{-1/2} \otimes I_T\right)y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\beta + \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\beta^{\text{ols}} - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\beta^{\text{ols}} \\
&= \left(\Omega^{-1/2} \otimes I_T\right)y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\beta^{\text{ols}} + \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\left(\beta^{\text{ols}} - \beta\right) \\
&= \underbrace{\left(\Omega^{-1/2} \otimes I_T\right)y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\beta^{\text{ols}}}_{=W} + \underbrace{\left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)\left(\beta^{\text{ols}} - \beta\right)}_{=Z} \\
&= W + Z
\end{aligned}$$

where β^{ols} is the OLS (and MLE) estimator given in (1.28). Hence, we have

$$\begin{aligned}
(y - (I_n \otimes \boldsymbol{X})\beta)'\left(\Omega^{-1} \otimes I_T\right)(y - (I_n \otimes \boldsymbol{X})\beta) &= (W + Z)'(W + Z) \\
&= W'W + 2Z'W + Z'Z.
\end{aligned}$$

We now consider each of the three terms separately.

Consider first $Z'Z$:

$$\begin{aligned} Z'Z &= \left(\beta^{\text{ols}} - \beta\right)' \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)' \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right) \left(\beta^{\text{ols}} - \beta\right) \\ &= \left(\beta - \beta^{\text{ols}}\right)' \left(\Omega^{-1/2} \otimes \boldsymbol{X}'\right) \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right) \left(\beta - \beta^{\text{ols}}\right) \\ &= \left(\beta - \beta^{\text{ols}}\right)' \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right) \left(\beta - \beta^{\text{ols}}\right). \end{aligned} \tag{1.29}$$

Consider next $W'W$:

$$\begin{aligned} W'W &= \left[\left(\Omega^{-1/2} \otimes I_T\right) y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right) \beta^{\text{ols}}\right]' \left[\left(\Omega^{-1/2} \otimes I_T\right) y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right) \beta^{\text{ols}}\right] \\ &= \left[\left(\Omega^{-1/2} \otimes I_T\right) \left(y - (I_n \otimes \boldsymbol{X})\, \beta^{\text{ols}}\right)\right]' \left[\left(\Omega^{-1/2} \otimes I_T\right) \left(y - (I_n \otimes \boldsymbol{X})\, \beta^{\text{ols}}\right)\right] \\ &= \left(y - (I_n \otimes \boldsymbol{X})\, \beta^{\text{ols}}\right)' \left(\Omega^{-1/2} \otimes I_T\right) \left(\Omega^{-1/2} \otimes I_T\right) \left(y - (I_n \otimes \boldsymbol{X})\, \beta^{\text{ols}}\right) \\ &= \left(y - (I_n \otimes \boldsymbol{X})\, \beta^{\text{ols}}\right)' \left(\Omega^{-1} \otimes I_T\right) \left(y - (I_n \otimes \boldsymbol{X})\, \beta^{\text{ols}}\right). \end{aligned}$$

Note that this expression is identical to that leading to formula (1.24) but replacing β by β^{ols}, so the same formula applies and we have

$$\begin{aligned} W'W &= \left(y - (I_n \otimes \boldsymbol{X})\, \beta^{\text{ols}}\right)' \left(\Omega^{-1} \otimes I_T\right) \left(y - (I_n \otimes \boldsymbol{X})\, \beta^{\text{ols}}\right) \\ & \quad \operatorname{tr}\left(\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{\text{ols}}\right)' \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{\text{ols}}\right) \Omega^{-1}\right) \\ &= \operatorname{tr}\left(\boldsymbol{S}^{\text{ols}} \Omega^{-1}\right). \end{aligned} \tag{1.30}$$

where

$$\boldsymbol{S}^{\text{ols}} \equiv \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{\text{ols}}\right)' \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{\text{ols}}\right) \tag{1.31}$$

is the sum of squared residuals from the OLS estimation of equation (1.16).

Finally, we will show that $Z'W = 0$:

$$\begin{aligned} Z'W &= \left(\beta^{\text{ols}} - \beta\right)' \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right)' \left[\left(\Omega^{-1/2} \otimes I_T\right) y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right) \beta^{\text{ols}}\right] \\ &= \left(\beta^{\text{ols}} - \beta\right)' \left(\Omega^{-1/2} \otimes \boldsymbol{X}'\right) \left[\left(\Omega^{-1/2} \otimes I_T\right) y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right) \beta^{\text{ols}}\right] \\ &= \left(\beta^{\text{ols}} - \beta\right)' \left[\left(\Omega^{-1/2} \otimes \boldsymbol{X}'\right) \left(\Omega^{-1/2} \otimes I_T\right) y - \left(\Omega^{-1/2} \otimes \boldsymbol{X}'\right) \left(\Omega^{-1/2} \otimes \boldsymbol{X}\right) \beta^{\text{ols}}\right] \\ &= \left(\beta^{\text{ols}} - \beta\right)' \left[\left(\Omega^{-1/2}\Omega^{-1/2} \otimes \boldsymbol{X}'\right) y - \left(\Omega^{-1/2}\Omega^{-1/2} \otimes (\boldsymbol{X}'\boldsymbol{X})\right) \beta^{\text{ols}}\right] \\ &= \left(\beta^{\text{ols}} - \beta\right)' \left[\left(\Omega^{-1} \otimes \boldsymbol{X}'\right) y - \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right) \beta^{\text{ols}}\right] \\ &= \left(\beta^{\text{ols}} - \beta\right)' \left[\left(\Omega^{-1} \otimes \boldsymbol{X}'\right) y - \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right) \beta^{\text{mle}}\right], \end{aligned}$$

where the last equality uses that $\beta^{\text{mle}} = \beta^{\text{ols}}$. The MLE first order condition (1.26) then implies that the term in square brackets is zero, implying that

$$Z'W = 0. \tag{1.32}$$

Putting together these three results we obtain

$$\begin{aligned} & \left(y-\left(I_n \otimes \boldsymbol{X}\right) \beta\right)^{\prime}\left(\Omega^{-1} \otimes I_T\right)\left(y-\left(I_n \otimes \boldsymbol{X}\right) \beta\right) \\ = & Z^{\prime} Z+W^{\prime} W \\ = & \left(\beta-\beta^{\text{ols}}\right)^{\prime}\left(\Omega^{-1} \otimes\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)\right)\left(\beta-\beta^{\text{ols}}\right)+\operatorname{tr}\left(\boldsymbol{S}^{\text{ols}} \Omega^{-1}\right) . \end{aligned} \tag{1.33}$$

Therefore, we can write the likelihood function (1.20) as

$$L(\beta, \Omega \mid Y)=\frac{|\Omega|^{-T / 2}}{(2 \pi)^{n T / 2}} \exp \left[-\frac{1}{2}\left(\beta-\beta^{\text{ols}}\right)^{\prime}\left(\Omega^{-1} \otimes\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)\right)\left(\beta-\beta^{\text{ols}}\right)\right] \exp \left[-\frac{1}{2} \operatorname{tr}\left(\boldsymbol{S}^{\text{ols}} \Omega^{-1}\right)\right] . \tag{1.34}$$

This is a standard decomposition of the likelihood function. The first exponent is the kernel of a multivariate normal with mean β^{ols} and covariance $\Omega \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}$. The second exponent looks something like a Wishart distribution. Recall that $\boldsymbol{X}$ is $T \times k$ with $k = np+1$, $\boldsymbol{Y}$ is $T \times n$, so that $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{\text{ols}}$ is $T \times n$ so that $\boldsymbol{S}^{\text{ols}} = \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{\text{ols}}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{\text{ols}}\right)$ is $n \times n$.

The term $\exp\left[-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{S}^{\text{ols}}\Omega^{-1}\right)\right]$ looks like the kernel of the inverted Wishart distribution (1.5) with scale matrix $\Psi = \boldsymbol{S}^{\text{ols}}$ and random matrix $Z = \Omega$. So we need to keep track and fix some of the terms in (1.34). As mentioned before, the first exponent is the kernel of a multivariate normal with mean β^{ols} and covariance $\Omega \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}$, so in the likelihood we must have the term $\left|\Omega \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\right|^{-1/2}$ but in (1.34) we have $|\Omega|^{-T/2}$. Also, recall that $\boldsymbol{X}'\boldsymbol{X}$ (and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$) is a $k \times k$ matrix. Then, using the property of the determinant of a Kronecker product we have

$$\left|\Omega \otimes\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|=|\Omega|^k\left|\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|^n$$

So let's write

$$\begin{aligned} |\Omega|^{-\frac{T}{2}} & =|\Omega|^{-\frac{k}{2}}|\Omega|^{\frac{k-T}{2}} \\ & =|\Omega|^{-\frac{k}{2}}\left|\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|^{-\frac{n}{2}}\left|\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|^{\frac{n}{2}}|\Omega|^{-\frac{T-k}{2}} \\ & =\left|\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|^{\frac{n}{2}}|\Omega|^{-\frac{k}{2}}\left|\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|^{-\frac{n}{2}}|\Omega|^{-\frac{T-k}{2}} \\ & =\left|\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|^{\frac{n}{2}}\left|\Omega \otimes\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|^{-1 / 2}|\Omega|^{-\frac{T-k}{2}} . \end{aligned}$$

Also, since β is a $kn \times 1$ vector, in the multivariate normal we must have the term $(2\pi)^{kn/2}$ but we have $(2\pi)^{Tn/2}$. But this is easy

$$(2 \pi) T n / 2=(2 \pi)^{k n / 2}(2 \pi)^{(T-k) n / 2} .$$

With these adjustments we can write (1.34) as

$$\begin{aligned} L(\beta, \Omega \mid Y)= & \left|\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|^{\frac{n}{2}} \frac{\left|\Omega \otimes\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)^{-1}\right|^{-1 / 2}}{(2 \pi)^{k n / 2}} \exp \left[-\frac{1}{2}\left(\beta-\beta^{\text{ols}}\right)^{\prime}\left(\Omega^{-1} \otimes\left(\boldsymbol{X}^{\prime} \boldsymbol{X}\right)\right)\left(\beta-\beta^{\text{ols}}\right)\right] \\ & \times \frac{|\Omega|^{-\frac{T-k}{2}}}{2 \pi)^{(T-k) n / 2}} \exp \left[-\frac{1}{2} \operatorname{tr}\left(\boldsymbol{S}^{\text{ols}} \Omega^{-1}\right)\right] . \end{aligned}$$

Clearly,

$$\frac{\left|\Omega \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\right|^{-1/2}}{(2\pi)^{kn/2}} \exp\left[-\frac{1}{2}\left(\beta - \beta^{\text{ols}}\right)'\left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right)\left(\beta - \beta^{\text{ols}}\right)\right]$$

is the density of a multivariate normal distribution for $\beta \in R^{kn}$ with mean β^{ols} and covariance matrix $\Omega \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}$, so that $\beta \sim N\left(\beta^{\text{ols}}, \Omega \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$.

As for the second term, since $\boldsymbol{S}^{\text{ols}}$ is $n \times n$, we only need to make a small adjustment and write $T - k = (T - k - n - 1) + n + 1$

$$|\Omega|^{-\frac{T-k}{2}} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}^{\text{ols}}\Omega^{-1}\right)\right] = |\Omega|^{-\frac{(T-k-n-1)+n+1}{2}} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}^{\text{ols}}\Omega^{-1}\right)\right]$$

which is the kernel of an inverted Wishart distribution (1.5) for Ω with scale matrix $\Psi = \boldsymbol{S}^{\text{ols}}$ and $T - k - n - 1$ degrees of freedom.

All in all, we proved that the likelihood function is proportional to the product of a normal distribution for β given Ω ,$\beta \sim N\left(\beta^{\text{ols}}, \Omega \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$, and an inverse Wishart distribution for $\Omega \sim iW\left(\boldsymbol{S}^{\text{ols}}, T - k - n - 1\right)$, and hence the likelihood function can finally be written as

$$L(\beta, \Omega|Y) \propto N\left(\beta^{\text{ols}}, \Omega \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \times iW\left(\boldsymbol{S}^{\text{ols}}, T - k - n - 1\right). \tag{1.35}$$

Equation (1.35) is very important as it will allows us to appropriately choose conjugate priors to obtain conditional posteriors in closed form. As we will see below, a Normal-Wishart prior conjugates the two blocks of the likelihood function.

Priors for VARs

In this section we will consider useful priors for our VAR(p) (1.14). Namely, we consider

a) A Normal prior for β assuming Ω is known (Theil mixed estimator).

b) A noninformative prior for both β and Ω (Jeffreys prior).

c) A Normal prior for β and a non-informative prior for Ω.

d) A Normal prior for β and an independent Wishart for Ω^{-1}.

Case 1. Normal prior for β for a given Ω

In this case we assume that we know Ω. To implement the procedure, we set Ω to its OLS estimated value. Let's assume the following prior for β

$$\beta \sim N(\beta_0, V_0),$$

so that

$$p(\beta) \propto |V_0|^{-0.5} \exp\left(-\frac{1}{2}(\beta - \beta_0)' V_0^{-1} (\beta - \beta_0)\right).$$

Using equation (1.20) we have

$$L\left(\beta,\Omega|Y\right)=\frac{|\Omega|^{-T/2}}{(2\pi)^{nT/2}}\exp\left[-\frac{1}{2}\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)'\left(\Omega^{-1}\otimes I_T\right)\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)\right].$$

Hence the posterior satisfies

$$\begin{aligned}p\left(\beta|\Omega,Y\right) &\propto p\left(\beta\right)L\left(\beta,\Omega|Y\right)\\ &\propto \exp\left[-\frac{1}{2}\left(\beta-\beta_0\right)'V_0^{-1}\left(\beta-\beta_0\right)\right]\exp\left[-\frac{1}{2}\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)'\left(\Omega^{-1}\otimes I_T\right)\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)\right]\\ &\propto \exp\left[-\frac{1}{2}\left[\left(\beta-\beta_0\right)'V_0^{-1}\left(\beta-\beta_0\right)+\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)'\left(\Omega^{-1}\otimes I_T\right)\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)\right]\right].\end{aligned}$$

The term in the exponent contains a sum of two quadratic expressions in β which will also be another quadratic expression in β. Let's decompose the two terms. First,

$$\left(\beta-\beta_0\right)'V_0^{-1}\left(\beta-\beta_0\right)=\beta'V_0^{-1}\beta-2\beta'V_0^{-1}\beta_0+\beta_0'V_0^{-1}\beta_0.$$

Also, when we derived the MLE estimation of β we proved that

$$\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)'\left(\Omega^{-1}\otimes I_T\right)\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)=y'\left(\Omega^{-1}\otimes I_T\right)y-2y'\left(\Omega^{-1}\otimes\boldsymbol{X}\right)\beta+\beta'\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)\beta$$

Hence,

$$\begin{aligned}&(\beta-\beta_0)'V_0^{-1}(\beta-\beta_0)+\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)'\left(\Omega^{-1}\otimes I_T\right)\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)\\ =&\beta'V_0^{-1}\beta-2\beta'V_0^{-1}\beta+\beta_0'V_0^{-1}\beta_0+y'(\Omega^{-1}\otimes I_T)y-2y'(\Omega^{-1}\otimes\boldsymbol{X})+\beta'(\Omega^{-1}\otimes(\boldsymbol{X}'\boldsymbol{X}))\beta\\ =&\beta'\left[V_0^{-1}+\Omega^{-1}\otimes(\boldsymbol{X}'\boldsymbol{X})\right]\beta-2\left[\beta_0'V_0^{-1}+y'(\Omega^{-1}\otimes\boldsymbol{X})\right]\beta+\beta_0'V_0^{-1}\beta_0+y'(\Omega^{-1}\otimes I_T)y\end{aligned}\tag{1.36}$$

The last two terms of the sum are constants.

We now write these two expressions as a quadratic equation in β of the form

$$\left(\beta-\beta_1\right)V_1^{-1}\left(\beta-\beta_1\right)$$

where C is a constant. Note that

$$\left(\beta-\beta_1\right)V_1^{-1}\left(\beta-\beta_1\right)=\beta'V_1^{-1}\beta-2\beta_1'V_1^{-1}\beta+\beta_1'V_1^{-1}\beta_1.\tag{1.37}$$

Comparing (1.36) with (1.37) we can match coefficients (this is called completing the square). First, define

$$V_1^{-1}=V_0^{-1}+\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)$$

and

$$\beta_1'V_1^{-1}=\beta_0'V_0^{-1}+y'\left(\Omega^{-1}\otimes\boldsymbol{X}\right)\Rightarrow\beta_1'=\left[\beta_0'V_0^{-1}+y'\left(\Omega^{-1}\otimes\boldsymbol{X}\right)\right]V_1$$

or

$$\beta_1=V_1\left[V_0^{-1}\beta_0+\left(\Omega^{-1}\otimes\boldsymbol{X}'\right)y\right]$$

This implies that

$$\left(\beta-\beta_0\right)'V_0^{-1}\left(\beta-\beta_0\right)+\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)'\left(\Omega^{-1}\otimes I_T\right)\left(y-\left(I_n\otimes\boldsymbol{X}\right)\beta\right)=\left(\beta-\beta_1\right)V_1^{-1}\left(\beta-\beta_1\right)+C$$

where

$$V_1 = \left[V_0^{-1} + \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right)\right]^{-1} \quad (1.38)$$

$$\beta_1 = V_1 \left[V_0^{-1}\beta_0 + \left(\Omega^{-1} \otimes \boldsymbol{X}'\right) y\right] \quad (1.39)$$

and C is a constant (i.e. does not depend on β).

With all this algebra we conclude that the posterior density satisfies

$$p(\beta|\Omega, Y) \propto N(\beta_1, V_1). \quad (1.40)$$

But we can do a little more. Take the term $(\Omega^{-1} \otimes \boldsymbol{X}') y$ in (1.39) and recall that the OLS estimate satisfies

$$\left(\Omega^{-1} \otimes \boldsymbol{X}'\right) y = \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right) \beta^{\text{ols}}.$$

Hence we can write (1.39) as

$$\begin{aligned}
\beta_1 &= V_1 \left[V_0^{-1}\beta_0 + \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right) \beta^{\text{ols}}\right] \\
&= \left[V_0^{-1} + \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right)\right]^{-1} \left[V_0^{-1}\beta_0 + \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right) \beta^{\text{ols}}\right] \\
&= \left[V_0^{-1} + \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right)\right]^{-1} V_0^{-1}\beta_0 + \left[V_0^{-1} + \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right)\right]^{-1} \left(\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right) \beta^{\text{ols}} \\
&= W\beta_0 + (I - W)\beta^{\text{ols}},
\end{aligned}$$

where the weighting matrix $W = \left[V_0^{-1} + (\Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X}))\right]^{-1} V_0^{-1}$. This implies that the posterior is a weighted average of the prior β_0 and the OLS estimate β^{ols}.

To implement this procedure, we proceed as follows:

```
i) Choose the parameters of the prior distribution of $\beta : \beta_0$ and $V_0$.
ii) Set $\Omega = \Omega^{\text{ols}}$, the OLS estimate of the covariance matrix.
iii) Draw from a Normal distribution with mean $\beta_1$ and covariance matrix
```

$$\left[V_0^{-1} + \Omega^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\right]^{-1}.$$

Although simple, this procedure still requires us to choose a large number of parameters for β_0 (nk parameters) and for its covariance matrix, V_0 ($kn(kn-1)/2$ parameters). This is too much. The literature has considered several approaches to simplify these choices. One prior that is quite popular in applied work is the *Minnesota* or *Litterman* prior, which we discuss next.

The Minnesota Prior

The Minnesota, or Litterman, prior, is a special case of Case 1 in which the prior parameters β_0 and V_0 are functions of a small number of hyperparameters. The original proposal is typically used for variables in levels and shrinks the VAR estimates toward a random walk for each variable. When the variables are in growth rates, it is common to specify the prior of the parameters in β to have all mean zero. The Minnesota prior has been useful in forecasting persistent economic time series. The traditional implementation for persistence series is as follows:

a) For each equation of the VAR, set the prior mean of the first lag of the dependent variable to one and set the prior mean of all other slope coefficients to zero. That is, the prior is that each variable follows an independent random walk,

$$y_{i,t} = y_{i,t-1} + v_{it}.$$

b) Set the prior variance of the ij^{th} element of the matrix D_ℓ, denoted by $v^D_{ij,\ell}$ to

$$v^D_{ij,\ell} = \begin{cases} \frac{\lambda_1}{\ell^{\lambda_3}} & \text{if } i = j \\ \frac{\lambda_1 \lambda_2}{\ell^{\lambda_3}} \frac{\sigma_i^2}{\sigma_j^2} & \text{if } i \neq j \end{cases}.$$

c) Set the prior variances of the intercept c (or other exogenous variables) as

$$v_i^c = \lambda_1 \lambda_4.$$

where λ_4 is usually a large number.

That is, the assumption is that the prior covariance matrix V_0 is a diagonal matrix that depends only on 4 parameters: λ_1, λ_2, λ_3, and λ_4. In this matrix, the variance of the coefficient in the i-th variable's own lags are set to $\lambda_1/\ell^{\lambda_3}$. That is, λ_1 is the variance of the coefficient on the first own lag and λ_3 is a decay parameter that controls how fast the variance on higher own lags shrinks toward zero. For variables j other than the i-th, the parameter λ_2 controls the relative tightness in the other lags compared to the own lag, so that a smaller λ_2 increases the relative tightness of the other lags. The term σ_j^2/σ_i^2 is a scale factor to account for the different variances of the variables. σ_i^2 is the i-th diagonal element of Ω. Finally, λ_4 measures the tightness of the constant or other exogenous variables in the VAR. Since normally we do not want to impose tight priors on these parameters, we usually set λ_4 to a large number (infinity).

Example: Consider a VAR with two endogenous variables and two lags along with a constant,

$$y_t = c + D_1 y_{t-1} + D_2 y_{t-2} + v_t.$$

In this case, each equation involves $k = np + 1 = 2 \times 2 + 1 = 5$ coefficients to estimate, for a total of $kn = 10$ parameters in β. Recall that

$$\beta = \text{vec}\left(\begin{bmatrix} D_1' \\ D_2' \\ c' \end{bmatrix}\right) = \text{vec}\left(\begin{bmatrix} D_1(1,1) & D_1(2,1) \\ D_1(1,2) & D_1(2,2) \\ D_2(1,1) & D_2(2,1) \\ D_2(1,2) & D_2(2,2) \\ c_1 & c_2 \end{bmatrix}\right) = \begin{bmatrix} D_1(1,1) \\ D_1(1,2) \\ D_2(1,1) \\ D_2(1,2) \\ c_1 \\ D_1(2,1) \\ D_1(2,2) \\ D_2(2,1) \\ D_2(2,2) \\ c_2 \end{bmatrix}$$

We set the Minnesota prior for the mean as

$$\beta_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} D_1(1,1) \\ D_1(1,2) \\ D_2(1,1) \\ D_2(1,2) \\ c_1 \\ D_1(2,1) \\ D_1(2,2) \\ D_2(2,1) \\ D_2(2,2) \\ c_2 \end{bmatrix}$$

The prior variance of β is set to

$$V_0 = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1\lambda_2\left(\frac{\sigma_1}{\sigma_2}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\lambda_1}{2^{\lambda_3}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\lambda_1\lambda_2}{2^{\lambda_3}}\left(\frac{\sigma_1}{\sigma_2}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_1\lambda_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_1\lambda_2\left(\frac{\sigma_2}{\sigma_1}\right)^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\lambda_1\lambda_2}{2^{\lambda_3}}\left(\frac{\sigma_2}{\sigma_1}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\lambda_1}{2^{\lambda_3}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1\lambda_4 \end{bmatrix}.$$

Case 2: Noninformative prior for β and Ω

Here we consider a "diffuse" prior for β. Jeffreys (1961) proposed a rule for generating priors that are noninformative about the parameter of interest and retain certain useful properties. The Jeffreys' (or diffuse) prior is proportional to the square root of the determinant of the Fisher information matrix. In the case of a VAR with n variables, the diffuse prior is

$$p(\beta,\Omega) = p(\beta)\,p(\Omega)$$

with

$$p(\beta) = \text{constant} \tag{1.41}$$

$$p(\beta,\Omega) \propto |\Omega|^{-(n+1)/2}. \tag{1.42}$$

Then, the posterior distribution is

$$p(\beta,\Omega|Y) = L(\beta,\Omega|Y)\,p(\beta)\,p(\Omega)$$

Write the likelihood as in (1.34)

$$L(\beta,\Omega|Y) \propto |\Omega|^{-T/2}\exp\left[-\frac{1}{2}\left(\beta-\beta^{\text{ols}}\right)'\left(\Omega^{-1}\otimes(\boldsymbol{X}'\boldsymbol{X})\right)\left(\beta-\beta^{\text{ols}}\right)\right]\times\exp\left[-\frac{1}{2}tr\left(\boldsymbol{S}^{\text{ols}}\Omega^{-1}\right)\right].$$

Above we showed that the likelihood can be written as

$$\begin{aligned}L(\beta,\Omega|Y) \propto &\left|\Omega\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(\beta-\beta^{\text{ols}}\right)'\left(\Omega\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)^{-1}\left(\beta-\beta^{\text{ols}}\right)\right]\\ &\times|\Omega|^{-\frac{(T-k-n-1)+n+1}{2}}\exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}^{\text{ols}}\Omega^{-1}\right)\right].\end{aligned}$$

Then the posterior is

$$\begin{aligned}p(\beta,\Omega|Y) \propto &\left|\Omega\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(\beta-\beta^{\text{ols}}\right)'\left(\Omega\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)^{-1}\left(\beta-\beta^{\text{ols}}\right)\right]\\ &\times|\Omega|^{-\frac{(T-k-n-1)+n+1}{2}}|\Omega|^{-(n+1)/2}\exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}^{\text{ols}}\Omega^{-1}\right)\right]\\ \propto &N\left(\beta^{\text{ols}},\Omega\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)|\Omega|^{-\frac{(T-k)+n+1}{2}}\exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}^{\text{ols}}\Omega^{-1}\right)\right]\\ \propto &N\left(\beta^{\text{ols}},\Omega\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)\times iW\left(\Omega|\boldsymbol{S}^{\text{ols}},T-k\right).\end{aligned}\tag{1.43}$$

Because the normal density integrates to 1, integrating over β we have

$$p\left(\Omega|Y\right)=\int_{\beta}p\left(\beta,\Omega|Y\right)d\beta=iW\left(\Omega|\boldsymbol{S}^{\text{ols}},T-k\right).$$

Using the definition of conditional probability

$$p\left(\beta,\Omega|Y\right)=p\left(\beta|\Omega,Y\right)P\left(\Omega|Y\right)$$

which implies, together with (1.43), that

$$p\left(\beta|\Omega,Y\right)=\frac{p\left(\beta,\Omega|Y\right)}{P\left(\Omega|Y\right)}\propto N\left(\beta^{\text{ols}},\Omega\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}\right).$$

Summarizing results, we have that

$$p\left(\beta|\Omega,Y\right)=N\left(\beta^{\text{ols}},\Omega\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}\right).\tag{1.44}$$

$$p\left(\Omega|Y\right)=iW\left(\Omega,\boldsymbol{S}^{\text{ols}},T-k\right)\tag{1.45}$$

It is possible to compute $p(\beta|Y)$ in closed form as done in Zellner (1971). But we can also simulate the posterior $p(\beta|Y)$ using a version of the Gibbs Sampler.

Algorithm (Gibbs Sampler): Choose a large N and set $j=1$. Then iterate on the following loop:

a) Draw Ω^j from $iW\left(\Omega|\boldsymbol{S}^{\text{ols}},T-k\right)$.

b) Draw β^j from $N(\beta^{\text{ols}}|\Omega^j\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1})$.

c) Store $\{\beta^j,\Omega^j\}$, set $j=j+1$, and return to step **a)** while $j<N$.

Case 3: Normal prior for β and non-informative prior for Ω

Here we consider the following prior

$$p\left(\beta,\Omega\right)=p\left(\beta\right)p\left(\Omega\right)$$

with

$$p\left(\beta\right)\sim N\left(\beta_0,V_0\right) \tag{1.46}$$

$$p\left(\Omega\right)\propto\left|\Omega\right|^{-(n+1)/2}. \tag{1.47}$$

As in the previous case, we will draw from the posterior using the Gibbs Sampler.

To that end, we first compute $p(\Omega|\beta,Y)$ assuming that β is known (i.e. conditioning on β). Using the likelihood representation (1.25)

$$\begin{aligned}L(\beta,\Omega|Y)=&\frac{\left|\Omega\right|^{-T/2}}{\left(2\pi\right)^{nT/2}}\exp\left[-\frac{1}{2}\text{tr}\left(\Omega^{-1}\boldsymbol{V}'\boldsymbol{V}\right)\right]\\ \propto&\left|\Omega\right|^{-T/2}\exp\left[-\frac{1}{2}\text{tr}\left(\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\right)\Omega^{-1}\right)\right]\end{aligned}$$

The joint posterior is then

$$\begin{aligned}p(\Omega|\beta,Y)\propto&\left|\Omega\right|^{-T/2}\exp\left[-\frac{1}{2}\text{tr}\left(\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\right)\Omega^{-1}\right)\right]\left|\Omega\right|^{-(n+1)/2}\\ \propto&\left|\Omega\right|^{-(T+n+1)/2}\exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}\Omega^{-1}\right)\right],\end{aligned}$$

where

$$\boldsymbol{S}=\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\right). \tag{1.48}$$

Therefore, the conditional distribution

$$p(\Omega|\beta,Y)=iW\left(\boldsymbol{S},T\right)$$

is an inverse Wishart with scale matrix $\boldsymbol{S}$ and T degrees of freedom.

Now suppose that Ω is known and, following the steps derived for Case 1, we conclude that

$$p\left(\beta|\Omega,Y\right)=N\left(\beta_1,V_1\right).$$

where

$$\begin{aligned}V_1=&\left[V_0^{-1}+\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)\right]^{-1}\\ \beta_1=&V_1\left[V_0^{-1}\beta_0+\left(\Omega^{-1}\otimes\left(\boldsymbol{X}'\boldsymbol{X}\right)\right)\beta^{\text{ols}}\right].\end{aligned}$$

Summarizing, we have the conditional posteriors $p(\Omega|\beta,Y)$ and $p(\beta|\Omega,Y)$, so we can use the Gibbs Sampler.

Algorithm (Gibbs Sampler): `Choose` N `large, an arbitrary` β^0`, and set` $j=1$`. Then iterate on the following loop:`

a) Draw Ω^j from the conditional posterior $iW(\boldsymbol{S}^{j-1}, T)$, where

$$\boldsymbol{S}^{j-1} = \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^{j-1}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^{j-1}\right)$$

is the sum of squared residuals using β^{j-1} as the VAR slope parameters.

b) Given Ω^j obtained in the previous step, draw $\beta^j \sim N(\beta_1, V_1)$ using the above formulas replacing Ω by Ω^j.

c) Set $j \to j+1$ and return to step a). Repeat many times and discard an initial burn in sample.

Case 4: The independent Normal-Wishart prior

This case is a Normal prior for β and inverse Wishart prior for Ω. The steps are quite similar to Case 3 with some minor modifications. Here we consider

$$p(\beta, \Omega) = p(\beta)p(\Omega)$$

with

$$p(\beta) \sim N(\beta_0, V_0) \propto \exp\left[-\frac{1}{2}(\beta - \beta^*)' V_\beta^{-1} (\beta - \beta^*)\right]$$

and

$$\Omega \sim iW(\boldsymbol{S}_0, v_0) \propto |\Omega|^{-\frac{v_0+n+1}{2}} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}_0\Omega^{-1}\right)\right].$$

As above, first compute $p(\Omega|\beta, Y)$ assuming that β is known by using the likelihood representation (1.25)

$$L(\beta, \Omega|Y) \propto |\Omega|^{-T/2} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}\Omega^{-1}\right)\right]$$

where $\boldsymbol{S}$ is given by (1.48). Then the conditional posterior is

$$\begin{aligned} p(\Omega|\beta, Y) &\propto |\Omega|^{-T/2} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}\Omega^{-1}\right)\right] |\Omega|^{-\frac{v^*+n+1}{2}} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{S}^*\Omega^{-1}\right)\right] \\ &\propto |\Omega|^{-(T+v_0+n+1)/2} \exp\left[-\frac{1}{2}\left[\text{tr}\left[\left(\boldsymbol{S} + \boldsymbol{S}_0\right)\Omega^{-1}\right]\right]\right] \\ &\propto iW\left(\boldsymbol{S} + \boldsymbol{S}_0, T + v_0\right). \end{aligned}$$

This is, of course, the kernel of an inverted Wishart distribution with $T + v_0$ degrees of freedom and $\boldsymbol{S} + \boldsymbol{S}_0$ scale coefficient matrix.

Now, conditional on Ω, the posterior of β we already know:

$$p(\beta|\Omega, Y) = N(\beta_1, V_1)$$

where

$$\begin{aligned} V_1 &= \left[V_0^{-1} + \left(\Omega^{-1} \otimes \boldsymbol{X}'\boldsymbol{X}\right)\right]^{-1} \\ \beta_1 &= V_1\left[V_0^{-1}\beta_0 + \left(\Omega^{-1} \otimes \boldsymbol{X}'\boldsymbol{X}\right)\beta^{\text{ols}}\right]. \end{aligned}$$

Summarizing, we have the conditional posteriors $p\left(\Omega|\beta, Y\right)$ and $p\left(\beta|\Omega, Y\right)$, so we can use the Gibbs Sampler. The algorithm is copied almost verbatim from that in the previous section, with the only difference in step 2

Algorithm: Choose a large N, an arbitrary β^0, and set $j = 1$. Then iterate on the following loop:

a) Draw Ω^j from the conditional posterior $iW(\boldsymbol{S}^{j-1} + \boldsymbol{S_0}, T + v_0)$, where

$$\boldsymbol{S}^{j-1} = \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B^{j-1}}\right)' \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B^{j-1}}\right)$$

is the sum of squared residuals using β^{j-1} at the VAR slope parameters.

b) Given Ω^j obtained in the previous step, draw $\beta^j \sim N(\beta_1, V_1)$ using the above formulas replacing Ω by Ω^j.

c) Set $j \to j+1$ and return to step a). Repeat many times and discard an initial burn in sample.