An Introduction to Bayesian Econometrics

Seth Leonard
seth@ottoquant.com

OttoQaunt
www.ottoquant.com

May 20, 2020

Bayesian econometrics is founded on Bayes' theorem, a simple equation which states that the probability of an event A conditional on B equals the probability of B conditional on A times the unconditional probability of A over the unconditional probability of B, that is,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

This result can be easily derived from the definition of conditional probability, which states that the probability of A conditional on B equals the probability of A and B occurring together over the probability of B, that is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2}$$

Using equation 2 we have

$$P(A|B)P(B) = P(A \cap B)$$

and

$$P(B|A)P(A) = P(A \cap B)$$

which, when combined, yields equation 1.

We can also write Bayes' theorem in terms of probability density functions (or probability mass functions for discrete distributions) by again using the definition of conditional probability

$$f(y|x) = \frac{f(x,y)}{f(x)} \tag{3}$$

That is, the distribution of y conditional on x equals the joint distribution of x and y over the distribution of x. The derivation of Bayes' theorem, which we use in estimating the distribution of parameters conditional on observed data (called the posterior distribution of the parameters), is the same as above:

$$f(y|x)f(x) = f(x,y)$$

and

$$f(x|y)f(y) = f(x,y)$$

thus

$$f(y|x) = \frac{f(x|y)f(y)}{f(x)} \tag{4}$$

In Bayesian jargon, $f(y)$ is our prior distribution (or just prior) for y and $f(y|x)$ is our posterior distribution (or just posterior) for y. The distribution $f(x|y)$ is the conditional distribution of x given y (for example the distribution of the data x conditional on the parameters y) and $f(x)$ is the marginal distribution of x, typically calculated as $f(x) = \int f(x,y)dy = \int f(x|y)f(y)dy$.

1 Basic Examples

1.1 Example 1: Testing for Illnesses

Testing for illnesses provides a morbid but illustrative example of Bayes rule. Suppose 1% of the population has an illness (but that there are no observable symptoms), that if an individual has the illness he always tests positive, and that the probability of a false positive is 10%. Denoting $A = 1$ as having the illness and $B = 1$ as testing positive then

$$A = \begin{cases} 1 & \textit{with prob. } 0.01 \\ 0 & \textit{with prob. } 0.99 \end{cases}$$

and

$$(B|A=1) = \begin{cases} 1 & \textit{with prob. } 1 \\ 0 & \textit{with prob. } 0 \end{cases} \qquad (B|A=0) = \begin{cases} 1 & \textit{with prob. } .1 \\ 0 & \textit{with prob. } .9 \end{cases}$$

Then

$$P(B=1|A=1)P(A=1) = 1 \times 0.01$$

and

$$\begin{aligned} P(B=1) &= P(B=1|A=1)P(A=1) + P(B=1|A=0)P(A=0) \\ &= 1 \times 0.01 + 0.1 \times 0.99 \\ &= 0.109 \end{aligned}$$

thus the probability that an individual who tests positive once is in fact ill is

$$\begin{aligned} P(A=1|B=1) &= \frac{P(B=1|A=1)P(A=1)}{P(B=1)} \\ &= \frac{0.01}{0.109} \\ &= 0.0917 \end{aligned}$$

If an individual tests positive twice, the probability that he is in fact ill is

$$\begin{aligned} P(A=1|B=1,1) &= \frac{P(B=1,1|A=1)P(A=1)}{P(B=1,1)} \\ &= \frac{0.01}{0.1\times0.1\times0.99+1\times1\times0.01} \\ &= 0.5025 \end{aligned}$$

1.2 Example 2: Normal data with a Bernoulli prior

This question was on Mark Watson's Gerzensee midterm.[1] Suppose some data Y follows a normal distribution with mean θ and standard deviation 1 and that θ follows a Bernoulli distribution with $p = 0.5$, that is, we know θ is either 1 or 0 and we attach equal probability to each outcome. Suppose we observe a single draw of $Y = 0$. What is our posterior for θ?

To answer this question we need the fact that

$$f_Y(Y|\theta) = \frac{1}{\sqrt{2\pi}} \exp\Big\{ -\frac{1}{2}(Y-\theta)^2 \Big\}$$

$$f_\theta(\theta) = \begin{cases} 0 & \textit{with prob. } 0.5 \\ 1 & \textit{with prob. } 0.5 \end{cases}$$

[1] Gerzensee is an economics program for first year Swiss PhDs.

Then evaluating $f_Y(0|\theta = 0) = \frac{1}{\sqrt{2\pi}}$, $f_Y(0|\theta = 1) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}$ we have that the denominator in Bayes rule for $Y = 0$ is

$$\begin{aligned} f_Y(0) &= f_Y(0|\theta = 0)f_\theta(0) + f_Y(0|\theta = 1)f_\theta(0) \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{2} + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}\frac{1}{2} \end{aligned}$$

The numerator (the conditional distribution times our prior) is, for $\theta = 0$ (recall θ can only take values 0 or 1)

$$f_Y(0|\theta = 0)f(\theta) = \frac{1}{\sqrt{2\pi}}\frac{1}{2}$$

thus for $\theta = 0$

$$f_\theta(0|Y = 0) = \frac{\frac{1}{\sqrt{2\pi}}\frac{1}{2}}{\frac{1}{\sqrt{2\pi}}\frac{1}{2} + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}\frac{1}{2}} = \frac{1}{1 + e^{-\frac{1}{2}}}$$

For $\theta = 1$

$$f_\theta(1|Y = 0) = \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}\frac{1}{2}}{\frac{1}{\sqrt{2\pi}}\frac{1}{2} + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}\frac{1}{2}} = \frac{e^{-\frac{1}{2}}}{1 + e^{-\frac{1}{2}}}$$

thus our posterior distribution for θ is

$$f(\theta|Y = 0) = \begin{cases} 0 & with\ prob.\ \frac{1}{1+e^{-\frac{1}{2}}} \\ 1 & with\ prob.\ \frac{e^{-\frac{1}{2}}}{1+e^{-\frac{1}{2}}} \end{cases}$$

1.3 Example 3: Normal-Normal Conjugate Density

Suppose we believe a parameter θ follows a normal distribution

$$f_p(\theta) = \frac{\lambda}{\sqrt{2\pi}} \exp\left\{-\frac{\lambda^2}{2}(\theta - \theta_0)^2\right\}$$

where θ_0 is our prior mean and $\frac{1}{\lambda}$ our prior standard deviation. Suppose also that we observe a vector X that follows a normal distribution

$$f_X(X|\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - \theta)^2\right\}$$

which we could alternatively write as

$$f_X(X|\theta) = (2\pi)^{-\frac{N}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X - \theta)'\Sigma^{-1}(X - \theta)\right\}$$

where $\Sigma = \sigma^2 I_N$ and N is the number of observations or elements of X (Note here that I have assumed the standard deviation of the data σ is known. This assumption facilitates deriving and computing the posterior for θ, but we will properly specify a prior for both θ and σ in section 2). To derive the posterior distribution of θ it is sufficient to show that this distribution also has the form of a normal; we need not worry about the constant $f_X(X)$ in the denominator of Bayes rule as this only ensures that the resulting distribution sums to one, that is, $\int_{-\infty}^{\infty} f_\theta(\theta|X)d\theta = 1$. Looking only at the numerator of Bayes rule, $f_X(X|\theta)f_p(\theta)$, we have

$$\begin{aligned} f_\theta(\theta|X) &\propto \exp\left\{-\tfrac{\lambda^2}{2}(\theta-\theta_0)^2\right\}\exp\left\{-\tfrac{1}{2\sigma^2}\textstyle\sum_i(x_i-\theta)^2\right\} \\ &\propto \exp\left\{-\tfrac{1}{2}\Big[\lambda^2(\theta^2-2\theta\theta_0+\theta_0^2)+\tfrac{1}{\sigma^2}(\textstyle\sum_i x_i^2-2\sum_i x_i\theta+N\theta^2)\Big]\right\} \end{aligned}$$

In the second term above the only parts we care about are those which relate to the parameter θ; the rest goes into the constant of integration which ensures $\int_{-\infty}^{\infty} f(\theta|X)d\theta = 1$, thus we can simply write the above term as

$$\begin{aligned} f_\theta(\theta|X) &\propto \exp\left\{-\tfrac{1}{2}\Big[\lambda^2(\theta^2-2\theta\theta_0)+\tfrac{1}{\sigma^2}(-2\textstyle\sum_i x_i\theta+N\theta^2)\Big]\right\} \\ &= \exp\left\{-\tfrac{1}{2\sigma^2}\Big[(\lambda\sigma)^2(\theta^2-2\theta\theta_0)-2\textstyle\sum_i x_i\theta+N\theta^2\Big]\right\} \end{aligned} \tag{5}$$

The trick now is to try and re-write equation 5 as a normal distribution. The easiest way to proceed is to guess and check (assuming, of course, that we can make a good guess of the posterior). To this end suppose

$$f_\theta(\theta|X) \propto \exp\left\{-\tfrac{N+(\lambda\sigma)^2}{2\sigma^2}\Big[\theta-\big(N+(\lambda\sigma)^2\big)^{-1}\big(\textstyle\sum_i x_i+(\lambda\sigma)^2\theta_0\big)\Big]^2\right\} \tag{6}$$

Multiplying out equation 6 we have

$$f_\theta(\theta|X) \propto \exp\left\{-\tfrac{1}{2\sigma^2}\Big[\big(N+(\lambda\sigma)^2\big)\theta^2-2\theta\big(\textstyle\sum_i x_i+(\lambda\sigma)^2\theta_0\big)+\frac{\sum_i x_i+\theta_0}{N+(\lambda\sigma)^2}\Big]\right\}$$

Again, since the term $\frac{\sum_i x_i+\theta_0}{N+(\lambda\sigma)^2}$ does not contain our parameter of interest θ we can dump it into the constant of integration and keep only the relevant part of the above term,

$$\begin{aligned} f_\theta(\theta|X) &\propto \exp\left\{-\tfrac{1}{2\sigma^2}\Big[\big(N+(\lambda\sigma)^2\big)\theta^2-2\theta\big(\textstyle\sum_i x_i+(\lambda\sigma)^2\theta_0\big)\Big]\right\} \\ &= \exp\left\{-\tfrac{1}{2\sigma^2}\Big[(\lambda\sigma)^2(\theta^2-2\theta\theta_0)-2\textstyle\sum_i x_i\theta+N\theta^2\Big]\right\} \end{aligned}$$

which is equation 5. Thus our posterior distribution of θ given the realization of the data X is also normal with mean $(N+\big(\lambda\sigma)^2\big)^{-1}\big(\sum_i x_i+(\lambda\sigma)^2\theta_0\big)$ and variance

$\frac{\sigma^2}{N+(\lambda\sigma)^2}$. Notice that as the standard deviation of our prior becomes smaller our posterior variance also shrinks (λ is therefore sometimes referred to as a "shrinkage parameter") but our posterior mean becomes biased. In the extreme case that $\lambda \to \infty$ our posterior mean is our prior mean θ_0 and our posterior variance is zero. Thus by specifying a prior we are reducing the variance of our parameter estimates at the cost of introducing a bias.

2 Bayesian Linear Regression

Bayesian Linear Regressions will become a sort of all purpose tool with which we will estimate both the observation equation and transition equation of our factor model when we come to estimation by simulation in section **??**. Our model in this section is a simple linear regression. In the univariate case we have

$$y = X\beta + \varepsilon \tag{7}$$

where y, the dependent variable, is a $T \times 1$ vector, X is a $T \times m$ matrix of explanatory variables, β is a $m \times 1$ vector of unknown parameters, and ε is a $T \times 1$ vector of shocks with distribution[2]

$$\varepsilon_t \sim \mathcal{N}(0, \sigma)$$

In the multivariate case our model is

$$Y = XB' + \varepsilon \tag{8}$$

in which Y is a $T \times k$ matrix of k dependent variables (observations t are indexed by row though the index here need not necessarily be time), X is a $T \times m$ matrix of m explanatory variables, B a $m \times k$ matrix of unknown parameters, and ε_t is a $T \times k$ matrix of shocks with distribution

$$\varepsilon_t \sim \mathcal{N}\Big([0], \Sigma\Big)$$

In each section below I begin with the derivation of the posterior when we assume the covariance of shocks is know before moving to normal-inverse gamma or normal-inverse Wishart conjugate priors which include posteriors for the covariance of shocks.

[2]I use σ here to denote variance *not* standard deviation as is often the case since for the multivariate model Σ also denotes variance.

2.1 Simple Univariate Linear Regression

Assuming initially that the distribution of shocks is known and beginning with the univariate case our prior for β is[3]

$$\pi(\beta) \propto \exp\left\{-\frac{1}{2\sigma}(\beta-\beta_0)'\Lambda_0(\beta-\beta_0)\right\}$$

where β_0 is our prior for β and Λ_0, our prior covariance of β, defines the strength of our prior beliefs. The distribution for our model in (7) is

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

so that our posterior is

$$f(\beta|y,X,\sigma) \propto \exp\left\{-\frac{1}{2\sigma}\left[(y-X\beta)'(y-X\beta)+(\beta-\beta_0)'\Lambda_0(\beta-\beta_0)\right]\right\}$$

Multiplying out the terms in square brackets above an ignoring those which do not contain β and thus become part of the constant of integration we have

$$f(\beta|y,X,\sigma) \propto -2y'X\beta+\beta'X'X\beta+\beta'\Lambda_0\beta-2\beta'\Lambda_0\beta_0 \tag{9}$$

Letting $\Lambda_n = (X'X+\Lambda_0)$ and again only keeping track of terms which do not form part of our constant of integration we have

$$\begin{aligned} f(\beta|y,X,\sigma) &\propto \exp\left\{-\tfrac{1}{2\sigma}\left[(\beta-\Lambda_n^{-1}(X'y+\Lambda_0\beta_0))'\Lambda_n(\beta-\Lambda_n^{-1}(X'y+\Lambda_0\beta_0))\right]\right\} \\ &\propto \exp\left\{-\tfrac{1}{2\sigma}\left[\beta'X'X\beta+\beta'\Lambda_0\beta-2\beta'X'y-2\beta'\Lambda_0\beta_0\right]\right\} \end{aligned}$$

These terms match those in equation (9), thus our posterior for β is distributed

$$\beta \sim \mathcal{N}\Big((X'X+\Lambda_0)^{-1}(X'y+\Lambda_0\beta_0), \sigma(X'X+\Lambda_0)^{-1}\Big) \tag{10}$$

This derivation assumes we know σ already, which of course will not be the case in practice.

[3]Specifying the variance of our prior in terms of the variance of shocks keeps the algebra much cleaner but is not strictly necessary. Also note that like Σ, σ denotes variance, not standard deviation as is often the case.

2.2 Simple Multivariate Linear Regression

To derive the posterior for our multivariate linear regression when we assume the distribution of shocks to the model $\varepsilon_t \sim \mathcal{N}\Big(0, \sigma\Big)$ is known begin by writing the model for an observation t as

$$y_t = Bx_t + \varepsilon_t$$

and define the vectorization of B as

$$\beta = vec(B)$$

Our prior for β is

$$\pi(\beta|\sigma) \propto \exp\Big\{ -\frac{1}{2}(\beta - \beta_0)'(V_0)^{-1}(\beta - \beta_0)\Big\}$$

where $V_0 = \Lambda^{-1} \otimes \sigma$ and Λ determines our prior tightness, that is, the strength of our prior beliefs. That is, a larger value for Λ corresponds to a small prior variance. The distribution of our model requires a few more definitions. Let y be the vector of observations y_t stacked over time (that is, $y = vec(Y')$) and $\mathbf{X}$ be the associated matrix of explanatory variables. Explicitly, $\mathbf{X} = X \otimes I_k$ where k is the number of observed variables in y_t. Then

$$f(y|\beta, \sigma, X) \propto \exp\Big\{-\frac{1}{2}(y - \mathbf{X}\beta)'\Sigma^{-1}(y - \mathbf{X}\beta)\Big\}$$

where $\Sigma^{-1} = I_T \otimes \sigma^{-1}$. Our posterior is then

$$f(\beta|X, y, \sigma) \propto \exp\Big\{ -\frac{1}{2}\Big[(\beta - \beta_0)'(V_0)^{-1}(\beta - \beta_0) + (y - \mathbf{X}\beta)'\Sigma^{-1}(y - \mathbf{X}\beta)\Big]\Big\}$$

Multiplying this expression out (and ignoring terms that don't contain β) leaves us with

$$\begin{aligned} f(\beta|X, y, \sigma) \quad &\propto \quad \exp\Big\{ -\tfrac{1}{2}\Big[\beta'(V_0)^{-1}\beta - 2y'\Sigma^{-1}\mathbf{X}\beta \\ &+ \quad \beta'\mathbf{X}'\Sigma^{-1}\mathbf{X}\beta - 2\beta'(V_0)^{-1}\beta_0\Big]\Big\} \end{aligned}$$

Using the property for Kronecker products that for conformable matrices $(A \otimes B)(C \otimes D) = AC \otimes BD$ this simplifies to

$$f(\beta|X, y, \sigma) \propto \exp\Big\{-\frac{1}{2}\Big[\beta'(V_0)^{-1}\beta - 2y'(X \otimes \sigma^{-1})\beta + \beta'(X'X \otimes \sigma^{-1}) - 2\beta'(V_0)^{-1}\beta_0\Big]\Big\}$$

in which case we can write the posterior as

$$f(\beta|X, y, \sigma) \propto \exp\Big\{-\frac{1}{2}\Big(\beta - V_1^{-1}\big((X' \otimes \sigma^{-1})y + V_0\beta_0\big)\Big)'V_1\Big(\beta - V_1^{-1}\big((X' \otimes \sigma^{-1})y + V_0\beta_0\big)\Big)\Big\}$$

where

$$V_1 = (X'X \otimes \sigma^{-1} + V_0^{-1})$$

or, using our definition of $V_0 = \Lambda^{-1} \otimes \sigma$,

$$V_1 = (X'X + \Lambda) \otimes \sigma^{-1}$$

That is, our posterior for β is normally distributed

$$\beta \sim \mathcal{N}\Big((X'X + \Lambda)^{-1}(X'y + \Lambda\beta_0), V_1^{-1}\Big)$$

In the future we can use a matrix normal distribution for our multivariate normal models which avoids the hassle of having to vectorize and then simplify our model, though it is hopefully useful to illustrate the vectorized version at least once.

2.3 Univariate Linear Regression with Normal-Inverse Gamma Prior

The previous two subsections have assumed the variances of shocks to our models are known. This simplifies the derivations but is not a very realistic assumption. These variances and, in the multivariate case, covariance matrices are particularly important when we come to filtering and smoothing using dynamic factor models as they determine how aggressively we should update our estimates of unobserved factors based on observables. Getting good estimates of the scale of shocks is therefore a high priority. Our conjugate prior (meaning the prior and posterior have the same form) for the univariate case models the parameters β as normally distributed while our prior for σ follows an inverse gamma distribution. Explicitly,

$$\pi(\beta|\sigma) \propto (\sigma)^{-\frac{k}{2}} \exp\left\{ -\frac{1}{2\sigma}(\beta - \beta_0)'\Lambda_0(\beta - \beta_0)\right\}$$

and

$$\pi(\sigma) \propto \sigma^{\frac{\nu_0}{2}-1} \exp\left\{-\frac{s_0}{2\sigma}\right\}$$

In the above inverse gamma distribution s_0 is our prior scale parameter. Loosely interpreted, this is our prior for the residual sum of squares in our model. ν_0 is our prior scale parameter, which we can interpret as the number of "prior observations". Increasing ν_0 will force our posterior for σ towards $\frac{1}{\nu_0}s_0$.

As we model innovations ε_t as normally distributed, our model has the form

$$f(y|\beta, \sigma, X) \propto \sigma^{-\frac{T}{2}} \exp\left\{-\frac{1}{2\sigma}(y - X\beta)'(y - X\beta)\right\}$$

so our posterior is

$$f(\beta,\sigma|y,X) \propto \underbrace{\sigma^{-\frac{T}{2}}\exp\left\{-\frac{1}{2\sigma}(y-X\beta)'(y-X\beta)\right\}}_{f(y|\beta,\sigma,X)}\times$$
$$\underbrace{(\sigma)^{-\frac{k}{2}}\exp\left\{-\frac{1}{2\sigma}(\beta-\beta_0)'\Lambda_0(\beta-\beta_0)\right\}}_{\pi(\beta|\sigma)}\underbrace{\sigma^{\frac{\nu_0}{2}-1}\exp\left\{-\frac{s_0}{2\sigma}\right\}}_{\pi(\sigma)}$$

Beginning by multiplying out the exponential terms we have

$$-\frac{1}{2\sigma}\Big[y'y-2y'X\beta+\beta'X'X\beta+\beta'\Lambda_0\beta-2\beta'\Lambda_0\beta_0+\beta_0'\Lambda_0\beta_0+s_0\Big] \tag{11}$$

If we define $\Lambda_T \equiv X'X+\Lambda_0$ then

$$\begin{array}{ll}\big(\beta-\Lambda_T^{-1}(X'y+\Lambda_0\beta_0)\big)'\Lambda_T\big(\beta-\Lambda_T^{-1}(X'y+\Lambda_0\beta_0) & = \\ \beta'\Lambda_T\beta-2\beta'(X'y+\Lambda_0\beta_0)+(X'y+\Lambda_0\beta_0)'\Lambda_T^{-1}(X'y+\Lambda_0\beta_0) & \end{array} \tag{12}$$

Comparing terms in 12 and 11 shows that we can re-write the exponential terms of the posterior as

$$\begin{array}{ll}-\frac{1}{2\sigma}\Big[\big(\beta-\Lambda_T^{-1}(X'y+\Lambda_0\beta_0)\big)'\Lambda_T\big(\beta-\Lambda_T^{-1}(X'y+\Lambda_0\beta_0)\big) & + \\ y'y-(X'y+\Lambda_0\beta_0)'\Lambda_T^{-1}(X'y+\Lambda_0\beta_0)+\beta_0'\Lambda_0\beta_0+s_0\Big] & \end{array} \tag{13}$$

The first line in 13 forms the normal part of our normal-inverse gamma posterior; the second line contains terms that will go into the posterior scale parameter of the inverse gamma distribution. We can re-write the scale parameter for the posterior by noting that the posterior mean for β is $\beta_T = \Lambda_T^{-1}(X'y+\Lambda_0\beta_0)\big)$. Thus

$$\begin{array}{ll}(y-X\beta_T)'(y-X\beta_T)+(\beta_T-\beta_0)'\Lambda_0(\beta_T-\beta_0) & = \\ y'y-2y'X\beta_T+\beta_TX'X\beta_T+\beta_T\Lambda_0\beta_T-2\beta_T'\Lambda_0\beta_0+\beta_0'\Lambda_0\beta_0 & = \\ y'y-2(y'X+\beta_0'\Lambda_0)\beta_T+\beta_T'\underbrace{(X'X+\Lambda_0)}_{\Lambda_T}\beta_T+\beta_0'\Lambda_0\beta_0 & \end{array} \tag{14}$$

Using the definitions of β_T and Λ_T

$$\beta_T'\Lambda_T\beta_T-(y'X+\beta_0'\Lambda_0)\beta_T=0$$

so that collecting all the terms in our posterior we have

$$\begin{array}{lll}f(\beta,\sigma|y,X) & \propto & \sigma^{-\frac{k}{2}}\exp\Big\{-\frac{1}{2\sigma}\big(\beta-\Lambda_T^{-1}(X'y+\Lambda_0\beta_0)\big)'\Lambda_T\big(\beta-\Lambda_T^{-1}(X'y+\Lambda_0\beta_0)\big)\Big\} \\ & \times & \sigma^{-\frac{\nu_0+T}{2}-1}\exp\Big\{-\frac{1}{2\sigma}\big((y-X\beta_T)'(y-X\beta_T)+(\beta_T-\beta_0)'\Lambda_0(\beta_T-\beta_0)+s_0\big)\Big\}\end{array}$$

That is, our posterior is a normal-inverse gamma distribution such that

$$f(\beta|\sigma, X, y) \sim \mathcal{N}\Big(\Lambda_T^{-1}(X'y + \Lambda_0\beta_0), \sigma\Lambda_T^{-1}\Big)$$

where $\Lambda_T = (X'X + \Lambda_0)$ and $f(\sigma|X, y)$ follows an inverse gamma distribution with scale parameter $s_T = (y - X\beta_T)'(y - X\beta_T) + (\beta_T - \beta_0)'\Lambda_0(\beta_T - \beta_0) + s_0$ and $\nu_T = T + \nu_0$.

2.4 Multivariate Linear Regression with Normal-Inverse Wishart Prior

To derive the posterior for our multivariate normal model

$$Y_t = B'X_t + \varepsilon_t$$

or, for all observations,

$$Y = XB + \varepsilon$$

we will use the multivariate equivalent of our normal-inverse gamma conjugate prior for the univariate case, that is, a matrix normal-inverse Wishart distribution.[4] Accordingly, our priors are

$$\pi(\Sigma) \sim \mathcal{IW}\big(V_0, \nu_0\big)$$

and

$$\pi(B|\Sigma) \sim \mathcal{MN}\big(B, \Lambda_0^{-1}, \Sigma\big)$$

which is equivalent to the vectorised form where $\beta = vec(B')$[5]

$$\pi(\beta|\Sigma) \sim \mathcal{N}(\beta_0, \Lambda_0^{-1} \otimes \Sigma)$$

Given these priors and the distribution for our model

$$f(Y|B, \Sigma, X) \sim \mathcal{MN}\big(XB, I_T, \Sigma\big)$$

[4]Note that in this section I will write B as its transpose for ease of notation. That is, B in this section corresponds to its transpose elsewhere. This is due to the fact that we will write our stacked model of observations as $Y = XB + \varepsilon$.

[5]I use the vectorized form $\beta = vec(B')$ as this stacks B over rows, thus the resulting vector β_0 corresponds to the covariance matrix $\Lambda_0^{-1} \otimes \Sigma$ and to $\mathbf{X} = X \otimes I_k$ as defined in the derivation for a simple normal-normal multivariate linear regression.

we can write our posterior as

$$\begin{aligned} f(B,\Sigma|Y,X) &\propto \underbrace{|\Sigma|^{-(\nu_0+k+1)/2}\exp\Big\{-\frac{1}{2}tr(V_0\Sigma^{-1})\Big\}}_{\pi(\Sigma)} \\ &\times \underbrace{|\Sigma|^{-m/2}\exp\Big\{-\frac{1}{2}tr\big((B-B_0)'\Lambda_0(B-B_0)\Sigma^{-1}\big)\Big\}}_{\pi(B|\Sigma)} \\ &\times \underbrace{|\Sigma|^{-T/2}\exp\Big\{-\frac{1}{2}tr\big((Y-XB)'(Y-XB)\Sigma^{-1}\big)\Big\}}_{f(Y|B,\Sigma,X)} \end{aligned}$$

As previously, the trick is to write our posterior as a sum of squares. Using the fact that $tr(A)tr(B) = tr(A+B)$ and dealing first with exponential terms, our result is the matrix equivalent to equation (11):

$$(15)\quad -\frac{1}{2}\Big[tr\Big(\big(Y'Y-2Y'XB+B'X'XB+B'\Lambda_0B-2B_0'\Lambda_0B+B_0'\Lambda_0B_0+V_0\big)\Sigma^{-1}\Big)\Big]$$

As previously, define $\Lambda_T = X'X + \Lambda_0$. We can again propose the sum of squares portion of our solution as

$$(16)\quad \begin{array}{l} \big(B-\Lambda_T^{-1}(X'Y+\Lambda_0B_0)\big)'\Lambda_T\big(B-\Lambda_T^{-1}(X'Y+\Lambda_0B_0)\big) \quad = \\ B'\Lambda_TB-2B'(X'Y+\Lambda_0B_0)+(X'Y+\Lambda_0B_0)'\Lambda_T^{-1}(X'Y+\Lambda_0B_0) \end{array}$$

Again, comparing terms in 16 and 15 shows that we can re-write the exponential terms of the posterior as

$$(17)\quad \begin{array}{l} -\frac{1}{2}tr\Bigg[\Big(\big(B-\Lambda_T^{-1}(X'Y+\Lambda_0B_0)\big)'\Lambda_T\big(B-\Lambda_T^{-1}(X'Y+\Lambda_0B_0)\big) \quad + \\ Y'Y-(X'Y+\Lambda_0B_0)'\Lambda_T^{-1}(X'Y+\Lambda_0B_0)+B_0'\Lambda_0B_0+V_0\Big)\Sigma^{-1}\Bigg] \end{array}$$

so that our final result is the matrix equivalent to our result in the univariate case, that is,

(18)

$$\begin{aligned} f(B,\Sigma|Y,X) &\propto |\Sigma|^{-m/2}\exp\Big\{-\tfrac{1}{2}tr\Big[\Big(\big(B-\Lambda_T^{-1}(X'Y+\Lambda_0B_0)\big)'\Lambda_T\big(B-\Lambda_T^{-1}(X'Y+\Lambda_0B_0)\big)\Big)\Sigma^{-1} \\ &\times \Sigma^{-\frac{1}{2}(\nu_0+k+T+1)}\Big\{tr\Big[\Big(\big((Y-XB_T)'(Y-XB_T)+(B_T-B_0)'\Lambda_0(B_T-B_0)+V_0\big)\Big) \end{aligned}$$

where $B_T = \Lambda_T^{-1}(X'Y+\Lambda_0B_0)$. Thus our posterior for B conditional on Σ and the data is

$$f(B|\Sigma,X,Y) \sim \mathcal{MN}\Big(\Lambda_T^{-1}(X'Y+\Lambda_0B_0),\Lambda_T,\Sigma\Big)$$

and our posterior for Σ conditional on the data is

$$f(\Sigma|X,Y) \sim \mathcal{IW}\Big((Y - XB_T)'(Y - XB_T) + (B_T - B_0)'\Lambda_0(B_T - B_0) + V_0, \nu_0 + T\Big)$$

where $\Lambda_T = X'X + \Lambda_0$ and $B_t = \Lambda_T^{-1}(X'Y + \Lambda_0 B_0)$. Note that for the practical purpose of drawing B in programs to simulate these posterior distributions we will need to use the vectorized form $\beta = vec(B')$ of

$$f(\beta|\Sigma,X,Y) \sim \mathcal{N}\Big(vec\big(\big[\Lambda_T^{-1}(X'Y + \Lambda B_0)\big]'\big), \Lambda_T^{-1} \otimes \Sigma\Big)$$

3 Estimation by Simulation

We can derive the posterior densities for the models in 2 by hand, making these models simple to estimate in practice. However, it will often be the case that we will not know the exact form of posterior densities. Dynamic factor models (DFMs) provide a good example. We can write our DFM as the observation equation

$$y_t = Hx_t + \varepsilon_t$$

and the transition equation

$$x_t = Ax_{t-1} + e_t$$

Assuming the distribution for shocks is

$$\begin{bmatrix} e_t \\ \varepsilon_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}\right)$$

If we knew the factors x_t, the posterior distributions for parameters H, R, A, Q follow from the results of section 2. However, x_t is itself an unobserved random variable. Thus we cannot calculate the posterior for H, R, A, Q and x_t directly. Instead, we will have to rely on simulation methods.

Markov chain Monte Carlo (MCMC) methods work by simulating a Markov chain that will have the same distribution as that which we would like to calculate, given enough iterations. That is, MCMC methods provide simulated values of distributions we cannot derive by hand. The following sections detail three of the most popular approaches to MCMC methods.

3.1 Metropolis Algorithm

Suppose we would like to estimate a density $P(x)$, but that $P(x)$ is difficult to work with directly. The Metropolis algorithm allows us to use a function $\pi(x)$ instead, where $\pi(x)$ is proportional to (but presumably simpler than) $P(x)$. The algorithm uses the following elements: $P(x)$ is the distribution we would like to simulate; we have a function from which we can draw values $\pi(x) \propto P(x)$; and we have a symmetric proposal distribution $g(x'|x)$ to generate proposed values of x for the next step of the iteration. The algorithm proceeds as follows:

1. Begin with an initial state x_0
2. Generate a new candidate x' from $g(x'|x_t)$. For example, $g(x'|x_t)$ might be $\mathcal{N}(x_t, 1)$. Using a normal distribution centered at x_t produces a random walk.
3. Calculate the acceptance ratio $a = \frac{\pi(x')}{\pi(x_t)} = \frac{P(x')}{P(x_t)}$. The last equality comes from the fact that $f(x) \propto P(x)$.
4. Accept x' with probability $\min\{1, a\}$. If x' is accepted, it becomes x_{t+1}. Otherwise, $x_{t+1} = x_t$.

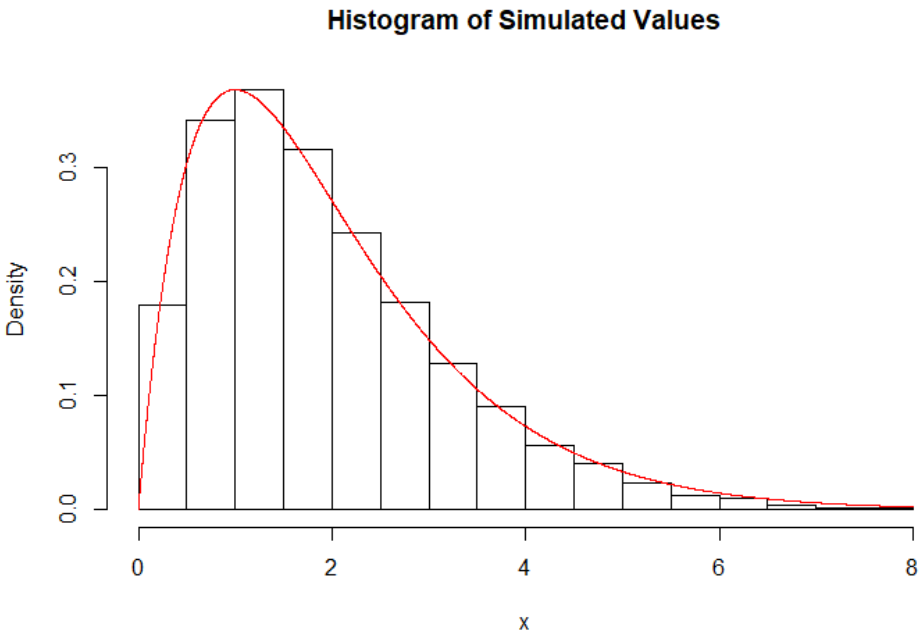
As an example, we'll use the gamma distribution

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \tag{19}$$

The fact that we can compute the actual distribution of (19) will be useful for testing our result. Because $\Gamma(\alpha)$ is not a function of x, we can use the proportional function

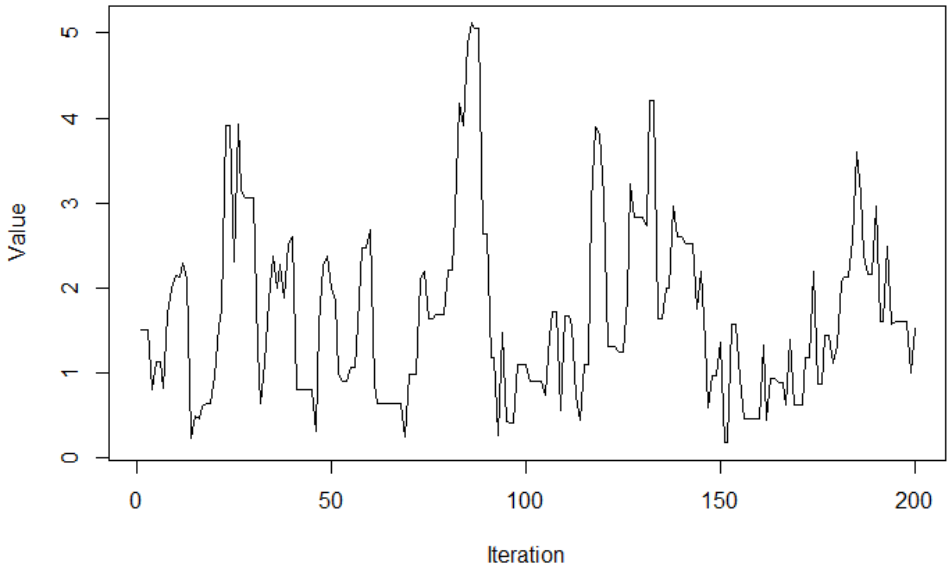
$$\pi(x) = \lambda e^{-\lambda x}(\lambda x)^{\alpha-1}$$

The following figure plots the resulting simulated distribution for 1000 burn in periods and 9000 iterations with shape parameter $\alpha = 2$ and scale parameter $\lambda = 1$.



Simulated distribution (histogram) versus true distribution (red)

When estimating distributions by simulations one might wish to get a sense of how well the simulation captures the desired distribution. A key concern is that the Markov chain we construct out of the above algorithm is stationary and mixes well. That is, are we consistently drawing from the same distribution at each iteration (a concern when using Gibbs sampling for factor models)? Second, do we explore the entire distribution, or does the algorithm get stuck in certain parts? The easiest way to asses this question is via a trace plot. A trace plot depicts draws, in this case for the variable x_t, at each iteration of the algorithm. What we wish to avoid is large discrete jumps in the algorithm, implying our Markov chain may not be stationary, and large flat sections, indicating the algorithm does not mix well. The following figure depicts the trace plot for x_t for the last 200 draws.



Trace plot for Metropolis simulations described in this section

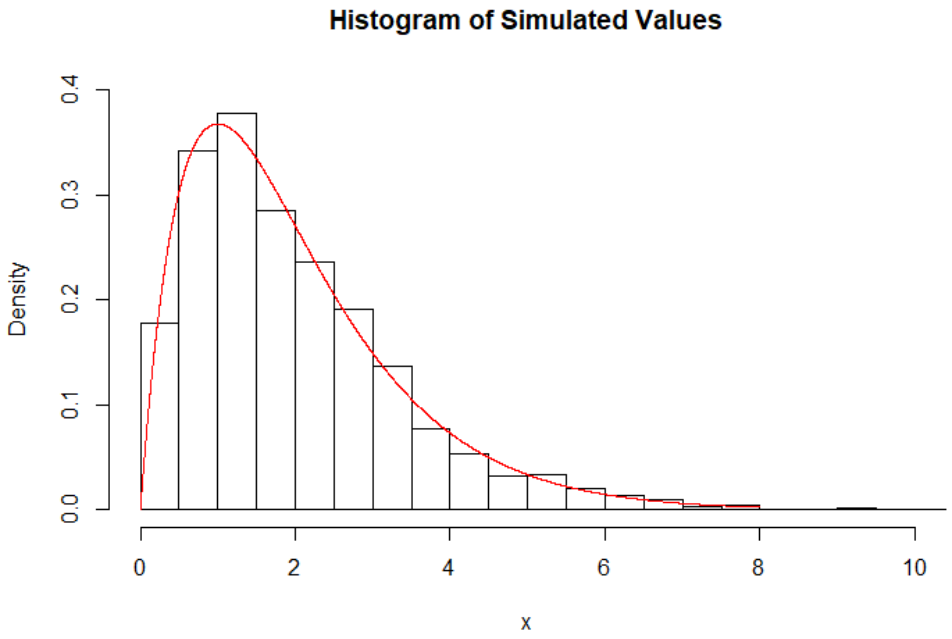
The results here look pretty good. There are some flat sections, but the algorithm seems to move around the distribution well.

3.2 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a generalization of the Metropolis algorithm which allows us to use a non-symmetric proposal distribution. Using the same notation as above, the algorithm proceeds as follows:

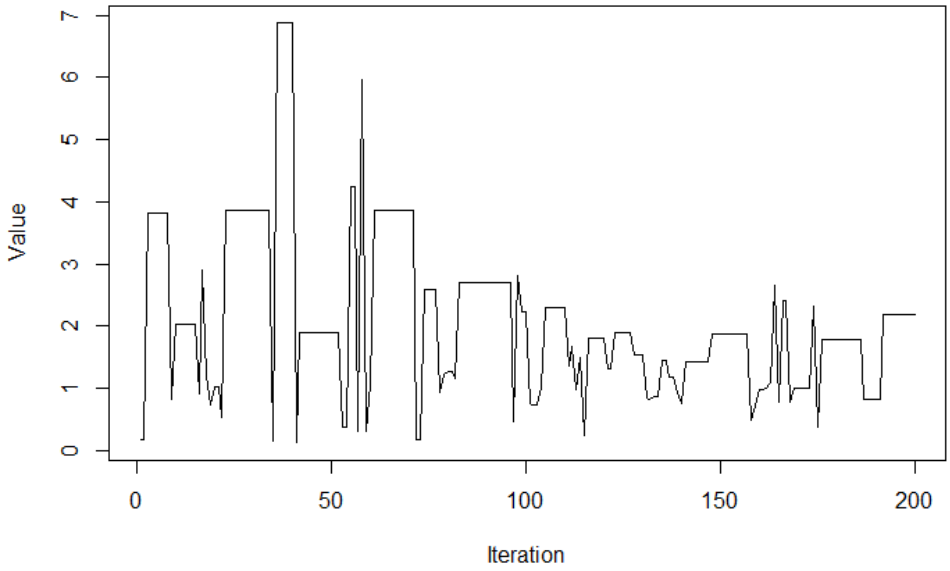
1. Begin with an initial state x_0
2. Generate a new candidate x' from $g(x'|x_t)$.
3. Calculate the acceptance ratio $a = \frac{\pi(x')g(x_t|x')}{\pi(x_t)g(x'|x_t)}$.
4. Accept x' with probability $\min\{1, a\}$. If x' is accepted, it becomes x_{t+1}. Otherwise, $x_{t+1} = x_t$.

We can again simulate our gamma distribution as a test case, this time using the non-symetric exponential distribution as our proposal distribution $g(x'|x_t) = x_t \exp^{-x_t x'}$. That is, we will use our current value x_t as the rate parameter. Our simulated distribution looks much the same as the previous example using the Metropolis algorithm.



Simulated distribution (histogram) versus true distribution (red)

However, looking at the trace plot for the last 200 draws reveals that this implementation of the Metropolis-Hastings algorithm did not mix as well as our previous implementation of the Metropolis algorithm.



Trace plot for Metropolis-Hastings simulations described in this section

3.3 Gibbs Sampling

Gibbs sampling is an MCMC method for simulating a multivariate distribution by iteratively sampling from conditional distributions. As we will see later, the approach is ideally suited to Bayesian estimation of dynamic factor models.

Suppose we want to sample $x = \begin{bmatrix} x_1 & x_2 & x_3 & \dots \end{bmatrix}$ from the distribution $p(x_1, x_2, x_3, \dots)$. We may not be able to draw from the joint distribution directly; instead we can draw from the conditional distributions $p(x_j | x_1, \dots, x_{j-1}, x_{j_1}, \dots)$. The algorithm proceeds as follows:

1. begin with an initial value $x^{(0)}$;
2. to get $x^{(1)}$, sample
   - $x_1^{(1)}$ from $p(x_1^{(1)} | x_2^{(0)}, x_2^{(0)}, x_3^{(0)}, \dots)$
   - $x_1^{(2)}$ from $p(x_1^{(2)} | x_1^{(1)}, x_3^{(0)}, x_4^{(0)}, \dots)$
   - $x_1^{(3)}$ from $p(x_1^{(3)} | x_1^{(1)}, x_2^{(1)}, x_4^{(0)}, \dots)$

and so on through all k elements of x;

3. repeat this process through the necessary burn in and sample periods.

Again it is instructive to take a simple example for which we know the true distribution. Suppose we were interested in the multivariate normal distribution

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \right) \tag{20}$$

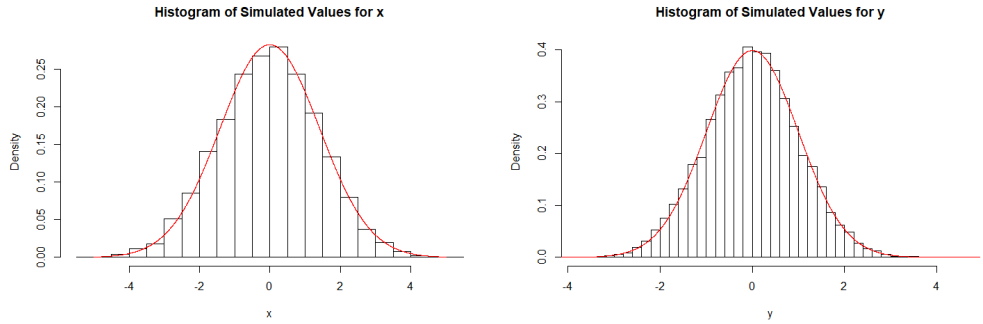
From the results we used to derive the Kalman filter, we know that

$$p(x_t|y_t) \sim \mathcal{N}\left(\frac{1}{2}y_t, \frac{7}{4}\right)$$

and

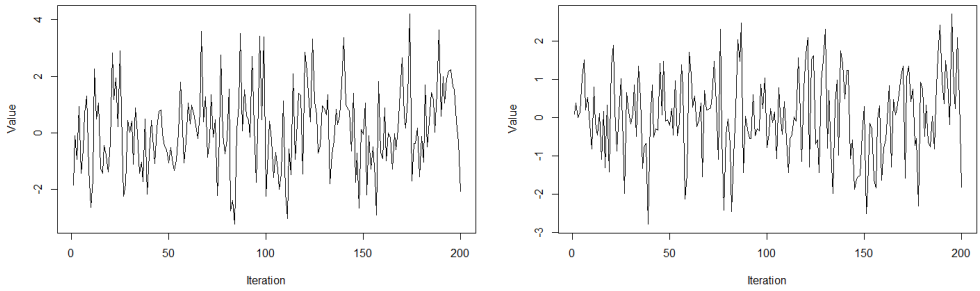
$$p(y_t|x_t) \sim \mathcal{N}\left(\frac{1}{4}x_t, \frac{7}{8}\right)$$

We can therefore simulate the joint distribution in (20) using these two conditional distributions via Gibbs sampling. The following figure illustrates the results for 1000 burn in periods and 9000 sampling periods.



Simulated distribution (histogram) for x (left) and y (right) versus true distribution (red)

Means of simulated values for x and y are -0.03 and -0.01 respectively, with variances 2.01 and 1.02. Trace plots, depicted below for the last 200 samples of x (left) and y (right) indicate that our Gibbs sampler mixes well.



Trace plots for Gibbs sampler depicting the last 200 draws for x (left) and y (right)

4 Approximating Bayesian Models

One great advantage of Bayesian econometric methods is that they allow us to reduce mean squared forecast error by introducing prior beliefs about parameters. Typically, our normally distributed priors will have a mean of zero. That is, we will begin with the prior that data is uninformative. While the results of section 2 provide closed for solutions for standard cases, we may not always want to use a formal Bayesian model. For example, we may wish to estimate a dynamic factor model by maximum likelihood. Maximum likelihood estimation has much weaker model identification requirements and avoids simulations, which may be computationally intensive. Inspection of our results in section 2 show us how we can "fake" a Bayesian prior in these cases, thereby benefiting from some of the advantages of Bayesian estimation without the complexity.

Recall from section 2 that we can write the posterior density for a normal-normal conjugate prior (assuming the variance Σ is known) where y_t is a scalar as

$$\beta \sim \mathcal{N}\Big((X'X + \Lambda_0)^{-1}(X'y + \Lambda_0\beta_0), \sigma(X'X + \Lambda_0)^{-1}\Big)$$

Suppose now that we use our standard prior of zero for β_0. Then our posterior simplifies to

$$\beta \sim \mathcal{N}\Big((X'X + \Lambda_0)^{-1}X'y, \sigma(X'X + \Lambda_0)^{-1}\Big)$$

Λ_0 is a symmetric (typically diagonal) matrix that determines the tightness of our prior. Thus, in order to calculate the posterior mean given a zero prior for β and prior tightness $\Lambda = I_k$, all we have to do is add an identity matrix to $X'X$ before taking its inverse. This simple trick is called a ridge regression. We could

even shrink estimates more aggressively towards zero by multiplying our identity matrix by a scalar λ. An additional advantage of a ridge regression (and Bayesian regression as well) is that it will allow us to estimate β even in cases where $X'X$ is singular (due to a lot of observations of zero, for example).

A second way to get at the same result is to add dummy data to X, with corresponding zero data in y. For example, suppose there were three series, $k = 3$, in the matrix X. We can tack a few identity matrices on to X to form a new matrix; for example, if we add three identity matrices we have

$$\tilde{X} = \begin{bmatrix} X \\ I_3 \\ I_3 \\ I_3 \end{bmatrix}$$

This gives us

$$\tilde{X}'\tilde{X} = X'X + \begin{bmatrix} I_3 & I_3 & I_3 \end{bmatrix} \begin{bmatrix} I_3 \\ I_3 \\ I_3 \end{bmatrix} = X'X + 3I_3$$

In this case we would also have to add nine zeros to the end of our series for y to ensure the dimensions agree. The result is the same as our posterior mean for β when $\Lambda_0 = 3I_3$. We could even include non-zero values for $\tilde{y}$. Suppose that we let

$$\tilde{y} = \begin{bmatrix} y \\ 3 \\ 2 \\ 1 \\ 3 \\ 2 \\ 1 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

Then

$$\tilde{X}'\tilde{y} = X'y + 3I_3 \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

The result is identical to the posterior mean for a model in which $\Lambda_0 = 3I_3$ and $\beta_0 = \begin{bmatrix} 3 & 2 & 1 \end{bmatrix}'$.

Both these methods for biasing estimates generalize immediately to multivariate models. Those these tricks sound very basic, they are a supremely useful tool in applied work, and have a clear Bayesian interpretation.

5 Summing Up

These results will form the basis of our Bayesian estimation routines for dynamic factor models. In later sessions, we will construct posterior densities for parameters of our model via Gibbs sampling. These simulations will use the conditional densities we have derived above. In summary, for a normal-inverse gamma conjugate prior our prior for β is

$$\pi(\beta|\sigma) \propto (\sigma)^{-\frac{k}{2}} \exp\left\{ -\frac{1}{2\sigma}(\beta-\beta_0)'\Lambda_0(\beta-\beta_0)\right\}$$

and our prior for sigma is

$$\pi(\sigma) \propto \sigma^{\frac{\nu_0}{2}-1} \exp\left\{-\frac{s_0}{2\sigma}\right\}$$

With the model distribution

$$f(y|\beta,\sigma,X) \propto \sigma^{-\frac{T}{2}} \exp\left\{-\frac{1}{2\sigma}(y-X\beta)'(y-X\beta)\right\}$$

the posterior distribution conditional on σ and the data X and y for β is

$$f(\beta|\sigma,X,y) \sim \mathcal{N}\Big(\Lambda_T^{-1}(X'y+\Lambda_0\beta_0), \sigma\Lambda_T^{-1}\Big)$$

where $\Lambda_T = (X'X+\Lambda_0)$ and $f(\sigma|X,y)$ follows an inverse gamma distribution with scale parameter $s_T = (y-X\beta_T)'(y-X\beta_T) + (\beta_T-\beta_0)'\Lambda_0(\beta_T-\beta_0) + s_0$ and $\nu_T = T+\nu_0$.

In the case of the multivariate model with a normal-inverse Wishart conjugate prior our prior for β conditional on σ is,[6] in vectorized form,

$$\pi(\beta|\sigma) \propto \exp\left\{ -\frac{1}{2}(\beta-\beta_0)'(\mathbf{\Lambda}_0)^{-1}(\beta-\beta_0)\right\}$$

[6]In using the vectorized form of the problem I use σ to denote the $k\times k$ covariance matrix of shocks ε_t and Σ to denote $I_T\otimes\sigma$. In using the matrix normal distribution Σ is sufficient to denote the $k\times k$ covariance matrix of shocks as using the Kronecker product is not necessary.

where $\mathbf{\Lambda}_0 = \Lambda_0 \otimes \sigma$ and Λ_0 determines our prior tightness. Using a prior scale parameter of I_k our prior for $\sigma \sim \mathcal{IW}(I_k, \nu_0)$ is

$$\pi(\sigma) \propto |\sigma|^{-\frac{\nu_0+k+1}{2}} \exp\left\{-\frac{1}{2} tr\big(\sigma^{-1}\big)\right\}$$

Our model is [7]

$$y_t = BX_t + \varepsilon_t$$

so that $\beta = vec(B)$ and thus the model distribution (again in vectorized form) is

$$f(y|\beta, \sigma, X) \propto \exp\left\{-\frac{1}{2}(y - \mathbf{X}\beta)'\Sigma^{-1}(y - \mathbf{X}\beta)\right\}$$

where $\mathbf{X} = X \otimes I_k$. Then the posterior for β conditional on σ is

$$f(\beta|\sigma, X, y) \sim \mathcal{N}\Big(vec\Big[\big(\Lambda_T^{-1}(X'Y + \Lambda_0 B_0')\big)'\Big], \Lambda_T^{-1} \otimes \sigma\Big)$$

where $\Lambda_T = X'X + \Lambda_0$ and $f(\sigma|X, y)$ follows an inverse-Wishart distribution with scale parameter $S_T = I_n + (Y - XB_T')'(Y - XB_T') + (B_T - B_0)\Lambda_0(B_T - B_0)'$ and $\nu_T = \nu_0 + T$. Note here that $B_T = \big(\Lambda_T^{-1}(X'Y + \Lambda_0 B_0')\big)'$ is the $k \times m$ matrix formed by stacking β_T and similarly B_0 is the stacked form of β_0.

6 Further Reading

These notes come largely from my own course notes from Gerzensee lectures with Mark Watson. The Wikipedia articles on Bayes' theorem and normal conjugate densities (including normal-inverse Wishart) are excellent. Sarkka (2013), freely available on-line, is also a good resource. Finally, Koop et al. (2007) discuss all the subjects covered in these notes in depth, and provide numerous solved problems and examples. For a much richer discussion of Monte Carlo methods see Robert and Casella (2010).

[7] Note here that B is back to its normal $k \times m$ orientation as opposed its definition in the previous section.

References

Koop, G., Poirier, D., and Tobias, J., 2007. *Bayesian Econometric Methods*. Cambridge University Press.

Robert, C. P. and Casella, G., 2010. *Monte Carlo Statistical Methods*. Springer Publishing Company, Incorporated.

Sarkka, S., 2013. *Bayesian Filtering and Smoothing*. Cambridge University Press.