12 Bayesian Analysis

The previous chapter noted that because so many parameters are estimated in a vector autoregression, the standard errors for inferences can be large. The estimates can be improved if the analyst has any information about the parameters beyond that contained in the sample. Bayesian estimation provides a convenient framework for incorporating prior information with as much weight as the analyst feels it merits.

Section 12.1 introduces the basic principles underlying Bayesian analysis and uses them to analyze a standard regression model or univariate autoregression. Vector autoregressions are discussed in Section 12.2. For the specifications in Sections 12.1 and 12.2, the Bayesian estimators can be found analytically. Numerical methods that can be used to analyze more general statistical problems from a Bayesian framework are reviewed in Section 12.3.

12.1. *Introduction to Bayesian Analysis*

Let $\boldsymbol{\theta}$ be an $(a \times 1)$ vector of parameters to be estimated from a sample of observations. For example, if $y_t \sim$ i.i.d. $N(\mu, \sigma^2)$, then $\boldsymbol{\theta} = (\mu, \sigma^2)'$ is to be estimated on the basis of $\mathbf{y} \equiv (y_1, y_2, \ldots, y_T)'$. Much of the discussion up to this point in the text has been based on the *classical* statistical perspective that there exists some true value of $\boldsymbol{\theta}$. This true value is regarded as an unknown but fixed number. An estimator $\hat{\boldsymbol{\theta}}$ is constructed from the data, and $\hat{\boldsymbol{\theta}}$ is therefore a random variable. In classical statistics, the mean and plim of the random variable $\hat{\boldsymbol{\theta}}$ are compared with the true value $\boldsymbol{\theta}$. The efficiency of the estimator is judged by the mean squared error of the random variable, $E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'$. A popular classical estimator is the value $\hat{\boldsymbol{\theta}}$ that maximizes the sample likelihood, which for this example would be

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - \mu)^2}{2\sigma^2}\right]. \qquad [12.1.1]$$

In Bayesian statistics, by contrast, $\boldsymbol{\theta}$ itself is regarded as a random variable. All inference about $\boldsymbol{\theta}$ takes the form of statements of probability, such as "there is only a 0.05 probability that θ_1 is greater than zero." The view is that the analyst will always have some uncertainty about $\boldsymbol{\theta}$, and the goal of statistical analysis is to describe this uncertainty in terms of a probability distribution. Any information the analyst had about $\boldsymbol{\theta}$ before observing the data is represented by a *prior density*

$f(\boldsymbol{\theta})$.[1] Probability statements that the analyst might have made about $\boldsymbol{\theta}$ before observing the data can be expressed as integrals of $f(\boldsymbol{\theta})$; for example, the previous statement would be expressed as $\int_0^{\infty} f(\theta_1)\, d\theta_1 = 0.05$ where $f(\theta_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\boldsymbol{\theta})\, d\theta_2\, d\theta_3 \cdots d\theta_n$. The sample likelihood [12.1.1] is viewed as the density of y conditional on the value of the random variable $\boldsymbol{\theta}$, denoted $f(\mathbf{y}|\boldsymbol{\theta})$. The product of the prior density and the sample likelihood gives the joint density of y and $\boldsymbol{\theta}$:

$$f(\mathbf{y}, \boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}). \qquad [12.1.2]$$

Probability statements that would be made about $\boldsymbol{\theta}$ after the data y have been observed are based on the *posterior density* of $\boldsymbol{\theta}$, which is given by

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}, \boldsymbol{\theta})}{f(\mathbf{y})}. \qquad [12.1.3]$$

Recalling [12.1.2] and the fact that $f(\mathbf{y}) = \int_{-\infty}^{\infty} f(\mathbf{y}, \boldsymbol{\theta})\, d\boldsymbol{\theta}$, equation [12.1.3] can be written as

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})}{\int_{-\infty}^{\infty} f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}, \qquad [12.1.4]$$

which is known as *Bayes's law*. In practice, the posterior density can sometimes be found simply by rearranging the elements in [12.1.2] as

$$f(\mathbf{y}, \boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{y}) \cdot f(\mathbf{y}),$$

where $f(\mathbf{y})$ is a density that does not involve $\boldsymbol{\theta}$; the other factor, $f(\boldsymbol{\theta}|\mathbf{y})$, is then the posterior density.

Estimating the Mean of a Gaussian Distribution with Known Variance

To illustrate the Bayesian approach, let $y_t \sim$ i.i.d. $N(\mu, \sigma^2)$ as before and write the sample likelihood [12.1.1] as

$$f(\mathbf{y}|\mu; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\mathbf{y} - \mu \cdot \mathbf{1})'(\mathbf{y} - \mu \cdot \mathbf{1})\right\}, \qquad [12.1.5]$$

where $\mathbf{1}$ denotes a $(T \times 1)$ vector of 1s. Here μ is regarded as a random variable. To keep the example simple, we will assume that the variance σ^2 is known with certainty. Suppose that prior information about μ is represented by the prior distribution $\mu \sim N(m, \sigma^2/\nu)$:

$$f(\mu; \sigma^2) = \frac{1}{(2\pi\sigma^2/\nu)^{1/2}} \exp\left[\frac{-(\mu - m)^2}{2\sigma^2/\nu}\right]. \qquad [12.1.6]$$

Here m and ν are parameters that describe the nature and quality of prior information about μ. The parameter m can be interpreted as the estimate of μ the analyst would have made before observing y, with σ^2/ν the *MSE* of this estimate. Expressing this *MSE* as a multiple $(1/\nu)$ of the variance of the distribution for y_t turns out to simplify some of the expressions that follow. Greater confidence in the prior information would be represented by larger values of ν.

[1]Throughout this chapter we will omit the subscript that indicates the random variable whose density is being described; for example, $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ will simply be denoted $f(\boldsymbol{\theta})$. The random variable whose density is being described should always be clear from the context and the argument of $f(\cdot)$.

To make the idea of a prior distribution more concrete, suppose that before observing y the analyst had earlier obtained a sample of N separate observations $\{z_i, i = 1, 2, \ldots, N\}$ from the $N(\mu, \sigma^2)$ distribution. It would then be natural to take m to be the mean of this earlier sample ($m = \bar{z} = (1/N)\sum_{i=1}^{N} z_i$) and σ^2/ν to be the variance of $\bar{z}$, that is, to take $\nu = N$. The larger this earlier sample (N), the greater the confidence in the prior information.

The posterior distribution for μ after observing the sample y is described by the following proposition.

Proposition 12.1: *The product of [12.1.5] and [12.1.6] can be written in the form* $f(\mu|\mathbf{y}; \sigma^2) \cdot f(\mathbf{y}; \sigma^2)$, *where*

$$f(\mu|\mathbf{y}; \sigma^2) = \frac{1}{[2\pi\sigma^2/(\nu + T)]^{1/2}} \exp\left[\frac{-(\mu - m^*)^2}{2\sigma^2/(\nu + T)}\right] \qquad [12.1.7]$$

$$f(\mathbf{y}; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{T/2}} |\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu|^{-1/2} \times \exp\left\{[-1/(2\sigma^2)](\mathbf{y} - m\cdot\mathbf{1})'(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu)^{-1}(\mathbf{y} - m\cdot\mathbf{1})\right\} \qquad [12.1.8]$$

$$m^* = \left(\frac{\nu}{\nu + T}\right)m + \left(\frac{T}{\nu + T}\right)\bar{y} \qquad [12.1.9]$$

$$\bar{y} = (1/T)\sum_{t=1}^{T} y_t.$$

In other words, the distribution of μ *conditional on the data* $(y_1, y_2, \ldots, y_T)$ *is* $N(m^*, \sigma^2/(\nu + T))$, *while the marginal distribution of* y *is* $N(m\cdot\mathbf{1}, \sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu))$.

With a quadratic loss function, the Bayesian estimate of μ is the value $\hat{\mu}$ that minimizes $E(\mu - \hat{\mu})^2$. Although this is the same expression as the classical *MSE*, its interpretation is different. From the Bayesian perspective, μ is a random variable with respect to whose distribution the expectation is taken, and $\hat{\mu}$ is a candidate value for the estimate. The optimal value for $\hat{\mu}$ is the mean of the posterior distribution described in Proposition 12.1:

$$\hat{\mu} = \left(\frac{\nu}{\nu + T}\right)m + \left(\frac{T}{\nu + T}\right)\bar{y}.$$

This is a weighted average of the estimate the classical statistician would use ($\bar{y}$) and an estimate based on prior information alone (m). Larger values of ν correspond to greater confidence in prior information, and this would make the Bayesian estimate closer to m. On the other hand, as ν approaches zero, the Bayesian estimate approaches the classical estimate $\bar{y}$. The limit of [12.1.6] as $\nu \to 0$ is known as a *diffuse* or *improper prior* density. In this case, the quality of prior information is so poor that prior information is completely disregarded in forming the estimate $\hat{\mu}$.

The uncertainty associated with the posterior estimate $\hat{\mu}$ is described by the variance of the posterior distribution. To use the data to evaluate the plausibility of the claim that $\mu_0 < \mu < \mu_1$, we simply calculate the probability $\int_{\mu_0}^{\mu_1} f(\mu|\mathbf{y}; \sigma^2)\, d\mu$. For example, the Bayesian would assert that the probability that μ is within the range $\hat{\mu} \pm 2\sigma/\sqrt{\nu + T}$ is 0.95.

Estimating the Coefficients of a Regression Model with Known Variance

Next, consider the linear regression model

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t,$$

where $u_t \sim$ i.i.d. $N(0, \sigma^2)$, $\mathbf{x}_t$ is a $(k \times 1)$ vector of exogenous explanatory variables, and $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of coefficients. Let

$$\underset{(T\times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} \qquad \underset{(T\times k)}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_T' \end{bmatrix}.$$

Treating $\boldsymbol{\beta}$ as random but σ^2 as known, we have the likelihood

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}; \sigma^2) &= \prod_{t=1}^{T} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](y_t - \mathbf{x}_t'\boldsymbol{\beta})^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \end{aligned} \qquad [12.1.10]$$

Suppose that prior information about $\boldsymbol{\beta}$ is represented by a $N(\mathbf{m}, \sigma^2\mathbf{M})$ distribution:

$$f(\boldsymbol{\beta}; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{k/2}} |\mathbf{M}|^{-1/2} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m})'\mathbf{M}^{-1}(\boldsymbol{\beta} - \mathbf{m})\right\}. \qquad [12.1.11]$$

Thus, prior to observation of the sample, the analyst's best guess as to the value of $\boldsymbol{\beta}$ is represented by the $(k \times 1)$ vector $\mathbf{m}$, and the confidence in this guess is summarized by the $(k \times k)$ matrix $\sigma^2\mathbf{M}$; less confidence is represented by larger diagonal elements of $\mathbf{M}$. Knowledge about the exogenous variables $\mathbf{X}$ is presumed to have no effect on the prior distribution, so that [12.1.11] also describes $f(\boldsymbol{\beta}|\mathbf{X}; \sigma^2)$.

Proposition 12.1 generalizes as follows.

Proposition 12.2: *The product of [12.1.10] and [12.1.11] can be written in the form* $f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}; \sigma^2) \cdot f(\mathbf{y}|\mathbf{X}; \sigma^2)$, *where*

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}; \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{k/2}} |\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X}|^{1/2} \\ &\quad \times \exp\left\{[-1/(2\sigma^2)](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \mathbf{m}^*)\right\} \end{aligned} \qquad [12.1.12]$$

$$\begin{aligned} f(\mathbf{y}|\mathbf{X}; \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{T/2}} |\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'|^{-1/2} \\ &\quad \times \exp\left\{[-1/(2\sigma^2)](\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m})\right\} \end{aligned} \qquad [12.1.13]$$

$$\mathbf{m}^* = (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{y}). \qquad [12.1.14]$$

In other words, the distribution of $\boldsymbol{\beta}$ *conditional on the observed data is* $N(\mathbf{m}^*, \sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1})$ *and the marginal distribution of* $\mathbf{y}$ *given* $\mathbf{X}$ *is* $N(\mathbf{X}\mathbf{m}, \sigma^2(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'))$.

Poor prior information about $\boldsymbol{\beta}$ corresponds to a large variance $\mathbf{M}$, or equivalently a small value for $\mathbf{M}^{-1}$. The diffuse prior distribution for this problem is often represented by the limit as $\mathbf{M}^{-1} \to \mathbf{0}$, for which the posterior mean [12.1.14] becomes $\mathbf{m}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the *OLS* estimator. The variance of the posterior distribution becomes $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Thus, classical regression inference is reproduced as a special case of Bayesian inference with a diffuse prior distribution. At the other extreme, if $\mathbf{X}'\mathbf{X} = \mathbf{0}$, the sample contains no information about $\boldsymbol{\beta}$ and the posterior distribution is $N(\mathbf{m}, \sigma^2\mathbf{M})$, the same as the prior distribution.

If the analyst's prior expectation is that all coefficients are zero ($\mathbf{m} = \mathbf{0}$) and this claim is made with the same confidence for each coefficient ($\mathbf{M}^{-1} = \lambda \cdot \mathbf{I}_k$ for some $\lambda > 0$), then the Bayesian estimator [12.1.14] is

$$\mathbf{m}^* = (\lambda \cdot \mathbf{I}_k + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \qquad [12.1.15]$$

which is the *ridge regression* estimator proposed by Hoerl and Kennard (1970). The effect of ridge regression is to shrink the parameter estimates toward zero.

Bayesian Estimation of a Regression Model with Unknown Variance

Propositions 12.1 and 12.2 assumed that the residual variance σ^2 was known with certainty. Usually, both σ^2 and $\boldsymbol{\beta}$ would be regarded as random variables, and Bayesian analysis requires a prior distribution for σ^2. A convenient prior distribution for this application is provided by the gamma distribution. Let $\{Z_i\}_{i=1}^N$ be a sequence of i.i.d. $N(0, \tau^2)$ variables. Then $W = \Sigma_{i=1}^N Z_i^2$ is said to have a gamma distribution with N degrees of freedom and scale parameter λ, indicated $W \sim \Gamma(N, \lambda)$, where $\lambda = 1/\tau^2$. Thus, W has the distribution of τ^2 times a $\chi^2(N)$ variable. The mean of W is given by

$$E(W) = N \cdot E(Z_i^2) = N\tau^2 = N/\lambda, \qquad [12.1.16]$$

and the variance is

$$\begin{aligned} E(W^2) - [E(W)]^2 &= N \cdot \{E(Z_i^4) - [E(Z_i^2)]^2\} \\ &= N \cdot (3\tau^4 - \tau^4) = 2N\tau^4 = 2N/\lambda^2. \end{aligned} \qquad [12.1.17]$$

The density of W takes the form

$$f(w) = \frac{(\lambda/2)^{N/2} w^{[(N/2)-1]} \exp[-\lambda w/2]}{\Gamma(N/2)}, \qquad [12.1.18]$$

where $\Gamma(\cdot)$ denotes the gamma function. If N is an even integer, then

$$\Gamma(N/2) = 1 \cdot 2 \cdot 3 \cdots [(N/2) - 1],$$

with $\Gamma(2/2) = 1$; whereas if N is an odd integer, then

$$\Gamma(N/2) = \sqrt{\pi} \cdot \tfrac{1}{2} \cdot \tfrac{3}{2} \cdot \tfrac{5}{2} \cdots [(N/2) - 1],$$

with $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Following DeGroot (1970) and Leamer (1978), it is convenient to describe the prior distribution not in terms of the variance σ^2 but rather in terms of the reciprocal of the variance, σ^{-2}, which is known as the *precision*. Thus, suppose that the prior distribution is specified as $\sigma^{-2} \sim \Gamma(N, \lambda)$, where N and λ are parameters that describe the analyst's prior information:

$$f(\sigma^{-2}|\mathbf{X}) = \frac{(\lambda/2)^{N/2} \sigma^{-2[(N/2)-1]} \exp[-\lambda\sigma^{-2}/2]}{\Gamma(N/2)}. \qquad [12.1.19]$$

Recalling [12.1.16], the ratio N/λ is the value expected for σ^{-2} on the basis of prior information. As we will see shortly in Proposition 12.3, if the prior information is based on an earlier sample of observations $\{z_1, z_2, \ldots, z_N\}$, the parameter N turns out to describe the size of this earlier sample and λ is the earlier sample's sum of squared residuals. For a given ratio of N/λ, larger values for N imply greater confidence in the prior information.

The prior distribution of $\boldsymbol{\beta}$ conditional on the value for σ^{-2} is the same as in [12.1.11]:

$$f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{k/2}} |\mathbf{M}|^{-1/2} \times \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m})'\mathbf{M}^{-1}(\boldsymbol{\beta} - \mathbf{m})\right\}. \quad [12.1.20]$$

Thus, $f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{X})$, the joint prior density for $\boldsymbol{\beta}$ and σ^{-2}, is given by the product of [12.1.19] and [12.1.20]. The posterior distribution $f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{y}, \mathbf{X})$ is described by the following proposition.

Proposition 12.3: *Let the prior density $f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{X})$ be given by the product of [12.1.19] and [12.1.20], and let the sample likelihood be*

$$f(\mathbf{y}|\boldsymbol{\beta}, \sigma^{-2}, \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \quad [12.1.21]$$

Then the following hold:

(*a*) *The joint posterior density of $\boldsymbol{\beta}$ and σ^{-2} is given by*

$$f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{y}, \mathbf{X}) = f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X}) \cdot f(\sigma^{-2}|\mathbf{y}, \mathbf{X}), \quad [12.1.22]$$

where the posterior distribution of $\boldsymbol{\beta}$ conditional on σ^{-2} is $N(\mathbf{m}^*, \sigma^2\mathbf{M}^*)$:

$$f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{k/2}} |\mathbf{M}^*|^{-1/2} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}, \quad [12.1.23]$$

with

$$\mathbf{m}^* = (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{y}) \quad [12.1.24]$$

$$\mathbf{M}^* = (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}. \quad [12.1.25]$$

Furthermore, the marginal posterior distribution of σ^{-2} is $\Gamma(N^*, \lambda^*)$:

$$f(\sigma^{-2}|\mathbf{y}, \mathbf{X}) = \frac{\sigma^{-2\{(N^*/2)-1\}}(\lambda^*/2)^{N^*/2}}{\Gamma(N^*/2)} \exp[-\lambda^*\sigma^{-2}/2], \quad [12.1.26]$$

with

$$N^* = N + T \quad [12.1.27]$$

$$\lambda^* = \lambda + (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \mathbf{m})'\mathbf{M}^{-1}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{m}) \quad [12.1.28]$$

for $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ the OLS estimator.

(*b*) *The marginal posterior distribution for $\boldsymbol{\beta}$ is a k-dimensional t distribution with N^* degrees of freedom, mean $\mathbf{m}^*$, and scale matrix $(\lambda^*/N^*) \cdot \mathbf{M}^*$:*

$$f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \left\{ \frac{\Gamma[(k + N^*)/2]}{(\pi N^*)^{k/2}\Gamma(N^*/2)} |(\lambda^*/N^*)\mathbf{M}^*|^{-1/2} \times [1 + (1/N^*)(\boldsymbol{\beta} - \mathbf{m}^*)'[(\lambda^*/N^*)\mathbf{M}^*]^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]^{-(k+N^*)/2} \right\}. \quad [12.1.29]$$

(c) *Let* $\mathbf{R}$ *be a known* $(m \times k)$ *matrix with linearly independent rows, and define*

$$Q \equiv \frac{[\mathbf{R}(\boldsymbol{\beta} - \mathbf{m}^*)]'[\mathbf{R}(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\cdot[\mathbf{R}(\boldsymbol{\beta} - \mathbf{m}^*)]/m}{\lambda^*/N^*}. \quad [12.1.30]$$

Then Q *has a marginal posterior distribution that is* $F(m, N^*)$:

$$f(q|\mathbf{y}, \mathbf{X}) = \frac{m^{m/2}(N^*)^{N^*/2}\Gamma[(N^* + m)/2]q^{[(m/2)-1]}}{\Gamma(m/2)\Gamma(N^*/2)(N^* + mq)^{[(N^*+m)/2]}}. \quad [12.1.31]$$

Recalling [12.1.16], result (a) implies that the Bayesian estimate of the precision is

$$E(\sigma^{-2}|\mathbf{y}, \mathbf{X}) = N^*/\lambda^*. \quad [12.1.32]$$

Diffuse prior information is sometimes represented as $N = \lambda = 0$ and $\mathbf{M}^{-1} = \mathbf{0}$. Substituting these values into [12.1.27] and [12.1.28] implies that $N^* = T$ and $\lambda^* = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$. For these values, the posterior mean [12.1.32] would be

$$E(\sigma^{-2}|\mathbf{y}, \mathbf{X}) = T/(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}),$$

which is the maximum likelihood estimate of σ^{-2}. This is the basis for the earlier claim that the parameter N for the prior distribution might be viewed as the number of presample observations on which the prior information is based and that λ might be viewed as the sum of squared residuals for these observations.

Result (b) implies that the Bayesian estimate of the coefficient vector is

$$E(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \mathbf{m}^* = (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{y}), \quad [12.1.33]$$

which is identical to the estimate derived in Proposition 12.2 for the case where σ^2 is known. Again, for diffuse prior information, $\mathbf{m}^* = \mathbf{b}$, the *OLS* estimate.

Result (c) describes the Bayesian perspective on a hypothesis about the value of $\mathbf{R}\boldsymbol{\beta}$, where the matrix $\mathbf{R}$ characterizes which linear combinations of the elements of $\boldsymbol{\beta}$ are of interest. A classical statistician would test the hypothesis that $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ by calculating an *OLS* F statistic,

$$\frac{(\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})/m}{s^2},$$

and evaluating the probability that an $F(m, T - k)$ variable could equal or exceed this magnitude. This represents the probability that the estimated value of $\mathbf{R}\mathbf{b}$ could be as far as it is observed to be from $\mathbf{r}$ given that the true value of $\boldsymbol{\beta}$ satisfies $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. By contrast, a Bayesian regards $\mathbf{R}\boldsymbol{\beta}$ as a random variable, the distribution for which is described in result (c). According to [12.1.30], the probability that $\mathbf{R}\boldsymbol{\beta}$ would equal $\mathbf{r}$ is related to the probability that an $F(m, N^*)$ variable would assume the value

$$\frac{(\mathbf{r} - \mathbf{R}\mathbf{m}^*)'[\mathbf{R}(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{m}^*)/m}{\lambda^*/N^*}.$$

The probability that an $F(m, N^*)$ variable could exceed this magnitude represents the probability that the random variable $\mathbf{R}\boldsymbol{\beta}$ might be as far from the posterior mean $\mathbf{Rm}^*$ as is represented by the point $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. In the case of a diffuse prior distribution, the preceding expression simplifies to

$$\frac{(\mathbf{r} - \mathbf{Rb})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb})/m}{(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})/T},$$

which is to be compared in this case with an $F(m, T)$ distribution. Recalling that

$$s^2 = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})/(T - k),$$

it appears that, apart from a minor difference in the denominator degrees of freedom, the classical statistician and the Bayesian with a diffuse prior distribution would essentially be calculating the identical test statistic and comparing it with the same critical value in evaluating the plausibility of the hypothesis represented by $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$.

Bayesian Analysis of Regressions with Lagged Dependent Variables

In describing the sample likelihood (expression [12.1.10] or [12.1.21]), the assumption was made that the vector of explanatory variables $\mathbf{x}_t$ was strictly exogenous. If $\mathbf{x}_t$ contains lagged values of y, then as long as we are willing to treat presample values of y as deterministic, the algebra goes through exactly the same. The only changes needed are some slight adjustments in notation and in the description of the results. For example, consider a pth-order autoregression with $\mathbf{x}_t = (1, y_{t-1}, y_{t-2}, \ldots, y_{t-p})'$. In this case, the expression on the right side of [12.1.21] describes the likelihood of $(y_1, y_2, \ldots, y_T)$ conditional on $y_0, y_{-1}, \ldots, y_{-p+1}$; that is, it describes $f(\mathbf{y}|\boldsymbol{\beta}, \sigma^{-2}, \mathbf{x}_1)$. The prior distributions [12.1.19] and [12.1.20] are then presumed to describe $f(\sigma^{-2}|\mathbf{x}_1)$ and $f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{x}_1)$, and the posterior distributions are all as stated in Proposition 12.3.

Note in particular that results (b) and (c) of Proposition 12.3 describe the exact small-sample posterior distributions, even when $\mathbf{x}_t$ contains lagged dependent variables. By contrast, a classical statistician would consider the usual t and F tests to be valid only asymptotically.

Calculation of the Posterior Distribution Using a GLS *Regression*

It is sometimes convenient to describe the prior information in terms of certain linear combinations of coefficients, such as

$$\mathbf{R}\boldsymbol{\beta}|\sigma^{-2} \sim N(\mathbf{r}, \sigma^2\mathbf{V}). \qquad [12.1.34]$$

Here $\mathbf{R}$ denotes a known nonsingular $(k \times k)$ matrix whose rows represent linear combinations of $\boldsymbol{\beta}$ in terms of which it is convenient to describe the analyst's prior information. For example, if the prior expectation is that $\beta_1 = \beta_2$, then the first row of $\mathbf{R}$ could be $(1, -1, 0, \ldots, 0)$ and the first element of $\mathbf{r}$ would be zero. The (1, 1) element of $\mathbf{V}$ reflects the uncertainty of this prior information. If $\boldsymbol{\beta} \sim N(\mathbf{m}, \sigma^2\mathbf{M})$, then $\mathbf{R}\boldsymbol{\beta} \sim N(\mathbf{Rm}, \sigma^2\mathbf{RMR}')$. Thus, the relation between the parameters for the prior distribution as expressed in [12.1.34] ($\mathbf{R}$, $\mathbf{r}$, and $\mathbf{V}$) and the parameters for the prior distribution as expressed in [12.1.20] ($\mathbf{m}$ and $\mathbf{M}$) is given by

$$\mathbf{r} = \mathbf{Rm} \qquad [12.1.35]$$

$$\mathbf{V} = \mathbf{RMR}'. \qquad [12.1.36]$$

Equation [12.1.36] implies

$$\mathbf{V}^{-1} = (\mathbf{R}')^{-1}\mathbf{M}^{-1}\mathbf{R}^{-1}. \qquad [12.1.37]$$

If equation [12.1.37] is premultiplied by $\mathbf{R}'$ and postmultiplied by $\mathbf{R}$, the result is

$$\mathbf{R}'\mathbf{V}^{-1}\mathbf{R} = \mathbf{M}^{-1}. \qquad [12.1.38]$$

Using equations [12.1.35] and [12.1.38], the posterior mean [12.1.33] can be rewritten as

$$\mathbf{m}^* = (\mathbf{R}'\mathbf{V}^{-1}\mathbf{R} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{R}'\mathbf{V}^{-1}\mathbf{r} + \mathbf{X}'\mathbf{y}). \qquad [12.1.39]$$

To obtain another perspective on [12.1.39], notice that the prior distribution [12.1.34] can be written

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad [12.1.40]$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$. This is of the same form as the observation equations of the regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad [12.1.41]$$

with $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$. The mixed estimation strategy described by Theil (1971, pp. 347–49) thus regards the prior information as a set of k additional observations, with r_i treated as if it were another observation on y, and the ith row of $\mathbf{R}$ corresponding to its vector of explanatory variables $\mathbf{x}_i'$. Specifically, equations [12.1.40] and [12.1.41] are stacked to form the system

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*, \qquad [12.1.42]$$

where

$$\underset{(T+k)\times 1}{\mathbf{y}^*} = \begin{bmatrix}\mathbf{r}\\ \mathbf{y}\end{bmatrix} \qquad \underset{(T+k)\times k}{\mathbf{X}^*} = \begin{bmatrix}\mathbf{R}\\ \mathbf{X}\end{bmatrix}$$

$$E(\mathbf{u}^*\mathbf{u}^{*\prime}) = \sigma^2\mathbf{V}^* = \sigma^2\begin{bmatrix}\mathbf{V} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_T\end{bmatrix}.$$

The *GLS* estimator for the stacked system is

$$\begin{aligned}\tilde{\mathbf{b}} &= [\mathbf{X}^{*\prime}(\mathbf{V}^*)^{-1}\mathbf{X}^*]^{-1}[\mathbf{X}^{*\prime}(\mathbf{V}^*)^{-1}\mathbf{y}^*]\\ &= \left\{[\mathbf{R}' \quad \mathbf{X}']\begin{bmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_T\end{bmatrix}\begin{bmatrix}\mathbf{R}\\ \mathbf{X}\end{bmatrix}\right\}^{-1} \times \left\{[\mathbf{R}' \quad \mathbf{X}']\begin{bmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_T\end{bmatrix}\begin{bmatrix}\mathbf{r}\\ \mathbf{y}\end{bmatrix}\right\}\\ &= (\mathbf{R}'\mathbf{V}^{-1}\mathbf{R} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{R}'\mathbf{V}^{-1}\mathbf{r} + \mathbf{X}'\mathbf{y}).\end{aligned}$$

Thus the posterior mean [12.1.39] can be calculated by *GLS* estimation of [12.1.42]. For known σ^2, the usual formula for the variance of the *GLS* estimator,

$$\sigma^2[\mathbf{X}^{*\prime}(\mathbf{V}^*)^{-1}\mathbf{X}^*]^{-1} = \sigma^2(\mathbf{R}'\mathbf{V}^{-1}\mathbf{R} + \mathbf{X}'\mathbf{X})^{-1},$$

gives a correct calculation of the variance of the Bayesian posterior distribution, $\sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$.

The foregoing discussion assumed that $\mathbf{R}$ was a nonsingular $(k \times k)$ matrix. On some occasions the analyst might have valuable information about some linear combinations of coefficients but not others. Thus, suppose that the prior distribution

[12.1.34] is written as

$$\begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{bmatrix} \boldsymbol{\beta} \sim N\left(\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix} \right),$$

where $\mathbf{R}_1$ is an $(m \times k)$ matrix consisting of those linear combinations for which the prior information is good and $\mathbf{R}_2$ is a $[(k - m) \times k]$ matrix of the remaining linear combinations. Then diffuse prior information about those linear combinations described by $\mathbf{R}_2$ could be represented by the limit as $\mathbf{V}_2^{-1} \to \mathbf{0}$, for which

$$\mathbf{R}'\mathbf{V}^{-1} = [\mathbf{R}_1' \quad \mathbf{R}_2'] \begin{bmatrix} \mathbf{V}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2^{-1} \end{bmatrix} \to [\mathbf{R}_1'\mathbf{V}_1^{-1} \quad \mathbf{0}].$$

The Bayesian estimate [12.1.39] then becomes

$$(\mathbf{R}_1'\mathbf{V}_1^{-1}\mathbf{R}_1 + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{R}_1'\mathbf{V}_1^{-1}\mathbf{r}_1 + \mathbf{X}'\mathbf{y}),$$

which can be calculated from *GLS* estimation of a $[(T + m) \times 1]$ system of the form of [12.1.42] in which only the linear combinations for which there is useful prior information are added as observations.

12.2. *Bayesian Analysis of Vector Autoregressions*

Litterman's Prior Distribution for Estimation of an Equation of a VAR

This section discusses prior information that might help improve the estimates of a single equation of a *VAR*. Much of the early econometric research with dynamic relations was concerned with estimation of distributed lag relations of the form

$$y_t = c + \omega_0 x_t + \omega_1 x_{t-1} + \cdots + \omega_p x_{t-p} + u_t. \qquad [12.2.1]$$

For this specification, ω_s has the interpretation as $\partial y_t/\partial x_{t-s}$, and some have argued that this should be a smooth function of s; see Almon (1965) and Shiller (1973) for examples. Whatever the merit of this view, it is hard to justify imposing a smoothness condition on the sequences $\{\omega_s\}_{s=1}^p$ or $\{\phi_s\}_{s=1}^p$ in a model with autoregressive terms such as

$$\begin{aligned} y_t = c &+ \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \\ &+ \omega_0 x_t + \omega_1 x_{t-1} + \cdots + \omega_p x_{t-p} + u_t, \end{aligned}$$

since here the dynamic multiplier $\partial y_t/\partial x_{t-s}$ is a complicated nonlinear function of the ϕ's and ω's.

Litterman (1986) suggested an alternative representation of prior information based on the belief that the change in the series is impossible to forecast:

$$y_t - y_{t-1} = c + \varepsilon_t, \qquad [12.2.2]$$

where ε_t is uncorrelated with lagged values of any variable. Economic theory predicts such behavior for many time series. For example, suppose that y_t is the log of the real price of some asset at time t, that is, the price adjusted for inflation. Then $y_t - y_{t-1}$ is approximately the real rate of return from buying the asset at $t - 1$ and selling it at t. In an extension of Fama's (1965) efficient markets argument described in Section 11.2, speculators would have bought more of the asset at time $t - 1$ if they had expected unusually high returns, driving y_{t-1} up in

relation to the anticipated value of y_t. The time path for $\{y_t\}$ that results from such speculation would exhibit price changes that are unforecastable. Thus, we might expect the real prices of items such as stocks, real estate, or precious metals to satisfy [12.2.2]. Hall (1978) argued that the level of spending by consumers should also satisfy [12.2.2], while Barro (1979) and Mankiw (1987) developed related arguments for the taxes levied and new money issued by the government. Changes in foreign exchange rates are argued by many to be unpredictable as well; see the evidence reviewed in Diebold and Nason (1990).

Write the ith equation in a *VAR* as

$$\begin{aligned} y_{it} = c_i &+ \phi_{i1}^{(1)} y_{1,t-1} + \phi_{i2}^{(1)} y_{2,t-1} + \cdots + \phi_{in}^{(1)} y_{n,t-1} \\ &+ \phi_{i1}^{(2)} y_{1,t-2} + \phi_{i2}^{(2)} y_{2,t-2} + \cdots + \phi_{in}^{(2)} y_{n,t-2} + \cdots \\ &+ \phi_{i1}^{(p)} y_{1,t-p} + \phi_{i2}^{(p)} y_{2,t-p} + \cdots + \phi_{in}^{(p)} y_{n,t-p} + \varepsilon_{it}, \end{aligned} \qquad [12.2.3]$$

where $\phi_{ij}^{(s)}$ gives the coefficient relating y_{it} to $y_{j,t-s}$. The restriction [12.2.2] requires $\phi_{ii}^{(1)} = 1$ and all other $\phi_{ij}^{(s)} = 0$. These values (0 or 1) then characterize the mean of the prior distribution for the coefficients. Litterman used a diffuse prior distribution for the constant term c_i.

Litterman took the variance-covariance matrix for the prior distribution to be diagonal, with γ denoting the standard deviation of the prior distribution for $\phi_{ii}^{(1)}$:

$$\phi_{ii}^{(1)} \sim N(1, \gamma^2).$$

Although each equation $i = 1, 2, \ldots, n$ of the *VAR* is estimated separately, typically the same number γ is used for each i. A smaller value for γ represents greater confidence in the prior information and will force the parameter estimates to be closer to the values predicted in [12.2.2]. A value of $\gamma = 0.20$ means that, before seeing the data, the analyst had 95% confidence that $\phi_{ii}^{(1)}$ is no smaller than 0.60 and no larger than 1.40.

The coefficients relating y_{it} to further lags are predicted to be zero, and Litterman argued that the analyst should have more confidence in this prediction the greater the lag. He therefore suggested taking $\phi_{ii}^{(2)} \sim N(0, (\gamma/2)^2)$, $\phi_{ii}^{(3)} \sim N(0, (\gamma/3)^2), \ldots,$ and $\phi_{ii}^{(p)} \sim N(0, (\gamma/p)^2)$, tightening the prior distribution with a harmonic series for the standard deviation as the lag increases.

Note that the coefficients $\phi_{ij}^{(s)}$ are scale-invariant; if each value of y_{it} is multiplied by 100, the values of $\phi_{ii}^{(s)}$ will be the same. The same is not true of $\phi_{ij}^{(s)}$ for $i \neq j$; if series i is multiplied by 100 but series j is not, then $\phi_{ij}^{(s)}$ will be multiplied by 100. Thus, in calculating the weight to be given the prior information about $\phi_{ij}^{(s)}$, an adjustment for the units in which the data are measured is necessary. Litterman proposed using the following standard deviation of the prior distribution for $\phi_{ij}^{(s)}$:

$$\frac{w \cdot \gamma \cdot \hat{\tau}_i}{s \cdot \hat{\tau}_j}. \qquad [12.2.4]$$

Here $(\hat{\tau}_i/\hat{\tau}_j)$ is a correction for the scale of series i compared with series j. Litterman suggested that $\hat{\tau}_i$ could be estimated from the standard deviation of the residuals from an *OLS* regression of y_{it} on a constant and on p of its own lagged values. Apart from this scale correction, [12.2.4] simply multiplies γ/s (which was the standard deviation for the prior distribution for $\phi_{ii}^{(s)}$) by a parameter w. Common experience with many time series is that the own lagged values $y_{i,t-s}$ are likely to

be of more help in forecasting y_{it} than will be values of other variables $y_{j,t-s}$. Hence we should have more confidence in the prior belief that $\phi_{ij}^{(s)} = 0$ than the prior belief that $\phi_{ii}^{(s)} = 0$, suggesting a value for w that is less than 1. Doan (1990) recommended a value of $w = 0.5$ in concert with $\gamma = 0.20$.

Several cautions in employing this prior distribution should be noted. First, for some series the natural prior expectation might be that the series is white noise rather than an autoregression with unit coefficient. For example, if y_{it} is a series such as the *change* in stock prices, then the mean of $\phi_{ii}^{(1)}$ should be 0 rather than 1. Second, many economic series display seasonal behavior. In such cases, $\phi_{ij}^{(s)}$ is likely to be nonzero for $s = 12$ and 24 with monthly data, for example. Litterman's prior distribution is not well suited for seasonal data, and some researchers suggest using seasonally adjusted data or including seasonal dummy variables in the regression before employing this prior distribution. Finally, the prior distribution is not well suited for systems that exhibit cointegration, a topic discussed in detail in Chapter 19.

Full-Information Bayesian Estimation of a VAR

Litterman's approach to Bayesian estimation of a *VAR* considered a single equation in isolation. It is possible to analyze all of the equations in a *VAR* together in a Bayesian framework, though the analytical results are somewhat more complicated than for the single-equation case; see Zellner (1971, Chapter 8) and Rothenberg (1973, pp. 139–44) for discussion.

12.3. *Numerical Bayesian Methods*

In the previous examples, the class of densities used to represent the prior information was carefully chosen in order to obtain a simple analytical characterization for the posterior distribution. For many specifications of interest, however, it may be impossible to find such a class, or the density that best reflects the analyst's prior information may not be possible to represent with this class. It is therefore useful to have computer-based methods to calculate or approximate posterior moments for a quite general class of problems.

Approximating the Posterior Mean by the Posterior Mode

One option is to use the mode rather than the mean of the posterior distribution, that is, to take the Bayesian estimate $\hat{\boldsymbol{\theta}}$ to be the value that maximizes $f(\boldsymbol{\theta}|\mathbf{y})$. For symmetric unimodal distributions, the mean and the mode will be the same, as turned out to be the case for the coefficient vector $\boldsymbol{\beta}$ in Proposition 12.2. Where the mean and mode differ, with a quadratic loss function the mode is a suboptimal estimator, though typically the posterior mode will approach the posterior mean as the sample size grows (see DeGroot, 1970, p. 236).

Recall from [12.1.2] and [12.1.3] that the posterior density is given by

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})}{f(\mathbf{y})}, \qquad [12.3.1]$$

and therefore the log of the posterior density is

$$\log f(\boldsymbol{\theta}|\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}) + \log f(\boldsymbol{\theta}) - \log f(\mathbf{y}). \qquad [12.3.2]$$

Note that if the goal is to maximize [12.3.2] with respect to $\boldsymbol{\theta}$, it is not necessary

to calculate $f(\mathbf{y})$, since this does not depend on $\boldsymbol{\theta}$. The posterior mode can thus be found by maximizing

$$\log f(\boldsymbol{\theta}, \mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}) + \log f(\boldsymbol{\theta}). \qquad [12.3.3]$$

To evaluate [12.3.2], we need only to be able to calculate the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ and the density that describes the prior information, $f(\boldsymbol{\theta})$. Expression [12.3.2] can be maximized by numerical methods, and often the same particular algorithms that maximize the log likelihood will also maximize [12.3.2]. For example, the log likelihood for a Gaussian regression model such as [12.1.21] can be maximized by a *GLS* regression, just as the posterior mode [12.1.39] can be calculated with a *GLS* regression.

Tierney and Kadane's Approximation for Posterior Moments

Alternatively, Tierney and Kadane (1986) noted that the curvature of the likelihood surface can be used to estimate the distance of the posterior mode from the posterior mean. Suppose that the objective is to calculate

$$E[g(\boldsymbol{\theta})|\mathbf{y}] = \int_{-\infty}^{\infty} g(\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}, \qquad [12.3.4]$$

where $\boldsymbol{\theta}$ is an $(a \times 1)$ vector of parameters and g: $\mathbb{R}^a \rightarrow \mathbb{R}^1$ is a function of interest. For example, if $g(\boldsymbol{\theta}) = \theta_1$, then [12.3.4] is the posterior mean of the first parameter, while $g(\boldsymbol{\theta}) = \theta_1^2$ gives the second moment. Expression [12.3.1] can be used to write [12.3.4] as

$$E[g(\boldsymbol{\theta})|\mathbf{y}] = \frac{\int_{-\infty}^{\infty} g(\boldsymbol{\theta}) \cdot f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{f(\mathbf{y})} = \frac{\int_{-\infty}^{\infty} g(\boldsymbol{\theta}) \cdot f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{-\infty}^{\infty} f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}. \qquad [12.3.5]$$

Define

$$h(\boldsymbol{\theta}) \equiv (1/T) \log\{g(\boldsymbol{\theta}) \cdot f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})\} \qquad [12.3.6]$$

and

$$k(\boldsymbol{\theta}) \equiv (1/T) \log\{f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})\}. \qquad [12.3.7]$$

This allows [12.3.5] to be written

$$E[g(\boldsymbol{\theta})|\mathbf{y}] = \frac{\int_{-\infty}^{\infty} \exp[T \cdot h(\boldsymbol{\theta})]\, d\boldsymbol{\theta}}{\int_{-\infty}^{\infty} \exp[T \cdot k(\boldsymbol{\theta})]\, d\boldsymbol{\theta}}. \qquad [12.3.8]$$

Let $\boldsymbol{\theta}^*$ be the value that maximizes [12.3.6], and consider a second-order Taylor series approximation to $h(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$:

$$\begin{aligned} h(\boldsymbol{\theta}) \cong h(\boldsymbol{\theta}^*) &+ \left.\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &+ \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)' \left\{\left.\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right\} (\boldsymbol{\theta} - \boldsymbol{\theta}^*). \end{aligned} \qquad [12.3.9]$$

Assuming that $\boldsymbol{\theta}^*$ is an interior optimum of $h(\cdot)$, the first derivative $[\partial h(\boldsymbol{\theta})/\partial \boldsymbol{\theta}']|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ is $\mathbf{0}$. Then [12.3.9] could be expressed as

$$h(\boldsymbol{\theta}) \cong h(\boldsymbol{\theta}^*) - (1/2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'(\boldsymbol{\Sigma}^*)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \tag{12.3.10}$$

where

$$\boldsymbol{\Sigma}^* \equiv -\left[\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right]^{-1}. \tag{12.3.11}$$

When [12.3.10] is substituted into the numerator of [12.3.8], the result is

$$\begin{aligned}
&\int_{-\infty}^{\infty} \exp[T \cdot h(\boldsymbol{\theta})]\, d\boldsymbol{\theta} \\
&\cong \int_{-\infty}^{\infty} \exp\left\{T \cdot h(\boldsymbol{\theta}^*) - (T/2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'(\boldsymbol{\Sigma}^*)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\} d\boldsymbol{\theta} \\
&= \exp[T \cdot h(\boldsymbol{\theta}^*)] \int_{-\infty}^{\infty} \exp\left\{(-T/2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'(\boldsymbol{\Sigma}^*)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\} d\boldsymbol{\theta} \\
&= \exp[T \cdot h(\boldsymbol{\theta}^*)]\, (2\pi)^{a/2} |\boldsymbol{\Sigma}^*/T|^{1/2} \\
&\quad \times \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{a/2}|\boldsymbol{\Sigma}^*/T|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'(\boldsymbol{\Sigma}^*/T)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\} d\boldsymbol{\theta} \\
&= \exp[T \cdot h(\boldsymbol{\theta}^*)]\, (2\pi)^{a/2} |\boldsymbol{\Sigma}^*/T|^{1/2}.
\end{aligned} \tag{12.3.12}$$

The last equality follows because the expression being integrated is a $N(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}^*/T)$ density and therefore integrates to unity.

Similarly, the function $k(\boldsymbol{\theta})$ can be approximated with an expansion around the posterior mode $\hat{\boldsymbol{\theta}}$,

$$k(\boldsymbol{\theta}) \cong k(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ maximizes [12.3.7] and

$$\hat{\boldsymbol{\Sigma}} \equiv -\left[\frac{\partial^2 k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^{-1}. \tag{12.3.13}$$

The denominator in [12.3.8] is then approximated by

$$\int_{-\infty}^{\infty} \exp[T \cdot k(\boldsymbol{\theta})]\, d\boldsymbol{\theta} \cong \exp[T \cdot k(\hat{\boldsymbol{\theta}})]\, (2\pi)^{a/2} |\hat{\boldsymbol{\Sigma}}/T|^{1/2}. \tag{12.3.14}$$

Tierney and Kadane's approximation is obtained by substituting [12.3.12] and [12.3.14] into [12.3.8]:

$$\begin{aligned}
E[g(\boldsymbol{\theta})|\mathbf{y}] &\cong \frac{\exp[T \cdot h(\boldsymbol{\theta}^*)]\, (2\pi)^{a/2} |\boldsymbol{\Sigma}^*/T|^{1/2}}{\exp[T \cdot k(\hat{\boldsymbol{\theta}})]\, (2\pi)^{a/2} |\hat{\boldsymbol{\Sigma}}/T|^{1/2}} \\
&= \frac{|\boldsymbol{\Sigma}^*|^{1/2}}{|\hat{\boldsymbol{\Sigma}}|^{1/2}} \exp\{T \cdot [h(\boldsymbol{\theta}^*) - k(\hat{\boldsymbol{\theta}})]\}.
\end{aligned} \tag{12.3.15}$$

To calculate this approximation to the posterior mean of $g(\boldsymbol{\theta})$, we first find the value $\boldsymbol{\theta}^*$ that maximizes $(1/T) \cdot \{\log g(\boldsymbol{\theta}) + \log f(\mathbf{y}|\boldsymbol{\theta}) + \log f(\boldsymbol{\theta})\}$. Then $h(\boldsymbol{\theta}^*)$ in [12.3.15] is the maximum value attained for this function and $\boldsymbol{\Sigma}^*$ is the negative of the inverse of the matrix of second derivatives of this function. Next we find the value $\hat{\boldsymbol{\theta}}$ that maximizes $(1/T) \cdot \{\log f(\mathbf{y}|\boldsymbol{\theta}) + \log f(\boldsymbol{\theta})\}$, with $k(\hat{\boldsymbol{\theta}})$ the maximum value attained and $\hat{\boldsymbol{\Sigma}}$ the negative of the inverse of the matrix of second derivatives.

The required maximization and second derivatives could be calculated analytically or numerically. Substituting the resulting values into [12.3.15] gives the Bayesian posterior estimate of $g(\boldsymbol{\theta})$.

Monte Carlo Estimation of Posterior Moments

Posterior moments can alternatively be estimated using the Monte Carlo approach suggested by Hammersley and Handscomb (1964, Section 5.4) and Kloek and van Dijk (1978). Again, the objective is taken to be calculation of the posterior mean of $g(\boldsymbol{\theta})$. Let $I(\boldsymbol{\theta})$ be some density function defined on $\boldsymbol{\theta}$ with $I(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta}$. Then [12.3.5] can be written

$$E[g(\boldsymbol{\theta})|\mathbf{y}] = \frac{\int_{-\infty}^{\infty} g(\boldsymbol{\theta}) \cdot f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{-\infty}^{\infty} f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}} = \frac{\int_{-\infty}^{\infty} \{g(\boldsymbol{\theta}) \cdot f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\} I(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{-\infty}^{\infty} \{f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\} I(\boldsymbol{\theta})\, d\boldsymbol{\theta}}. \quad [12.3.16]$$

The numerator in [12.3.16] can be interpreted as the expectation of the random variable $\{g(\boldsymbol{\theta}) \cdot f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\}$, where this expectation is taken with respect to the distribution implied by the density $I(\boldsymbol{\theta})$. If $I(\boldsymbol{\theta})$ is a known density such as multivariate Gaussian, it may be simple to generate N separate Monte Carlo draws from this distribution, denoted $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)}\}$. We can then calculate the average realized value of the random variable across these Monte Carlo draws:

$$\sum_{i=1}^{N} (1/N) \cdot \{g(\boldsymbol{\theta}^{(i)}) \cdot f(\mathbf{y}|\boldsymbol{\theta}^{(i)}) \cdot f(\boldsymbol{\theta}^{(i)})/I(\boldsymbol{\theta}^{(i)})\}. \quad [12.3.17]$$

From the law of large numbers, as $N \to \infty$, this will yield a consistent estimate of

$$E_{I(\boldsymbol{\theta})}\{g(\boldsymbol{\theta}) \cdot f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\} = \int_{-\infty}^{\infty} \{g(\boldsymbol{\theta}) \cdot f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\} I(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \quad [12.3.18]$$

provided that the integral in [12.3.18] exists. The denominator of [12.3.16] is similarly estimated from

$$\sum_{i=1}^{N} (1/N) \cdot \{f(\mathbf{y}|\boldsymbol{\theta}^{(i)}) \cdot f(\boldsymbol{\theta}^{(i)})/I(\boldsymbol{\theta}^{(i)})\}.$$

The integral in [12.3.18] need not exist if the importance density $I(\boldsymbol{\theta})$ goes to zero in the tails faster than the sample likelihood $f(\mathbf{y}|\boldsymbol{\theta})$. Even if [12.3.18] does exist, the Monte Carlo average [12.3.17] may give a poor estimate of [12.3.18] for moderate N if $I(\boldsymbol{\theta})$ is poorly chosen. Geweke (1989) provided advice on specifying $I(\boldsymbol{\theta})$. If the set of allowable values for $\boldsymbol{\theta}$ forms a compact set, then letting $I(\boldsymbol{\theta})$ be the density for the asymptotic distribution of the maximum likelihood estimator is usually a good approach.

A nice illustration of the versatility of Bayesian Monte Carlo methods for analyzing dynamic models is provided by Geweke (1988a). This approach was extended to multivariate dynamic systems in Geweke (1988b).

APPENDIX 12.A. *Proofs of Chapter 12 Propositions*

■ **Proof of Proposition 12.1.** Note that the product of [12.1.5] and [12.1.6] can be written

$$f(\mathbf{y}, \mu; \sigma^2) = \frac{1}{(2\pi)^{(T+1)/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right\}, \qquad [12.A.1]$$

where

$$\underset{(T+1)\times 1}{\boldsymbol{\alpha}} \equiv \begin{bmatrix} \mu - m \\ \mathbf{y} - \mu\cdot\mathbf{1} \end{bmatrix} \qquad [12.A.2]$$

$$\underset{(T+1)\times(T+1)}{\boldsymbol{\Sigma}} \equiv \begin{bmatrix} \sigma^2/\nu & \mathbf{0}' \\ \mathbf{0} & \sigma^2\mathbf{I}_T \end{bmatrix}. \qquad [12.A.3]$$

The goal is to rearrange $\boldsymbol{\alpha}$ so that μ appears only in the first element. Define

$$\underset{(T+1)\times(T+1)}{\mathbf{A}} \equiv \begin{bmatrix} \nu/(\nu + T) & -\mathbf{1}'/(\nu + T) \\ \mathbf{1} & \mathbf{I}_T \end{bmatrix}. \qquad [12.A.4]$$

Since $\mathbf{1}'\mathbf{1} = T$ and $\mathbf{1}'\mathbf{y} = T\bar{y}$, we have

$$\begin{aligned} \mathbf{A}\boldsymbol{\alpha} &= \begin{bmatrix} [\nu/(\nu + T)](\mu - m) - \mathbf{1}'\mathbf{y}/(\nu + T) + [T/(\nu + T)]\mu \\ \mathbf{y} - m\cdot\mathbf{1} \end{bmatrix} \\ &= \begin{bmatrix} \mu - m^* \\ \mathbf{y} - m\cdot\mathbf{1} \end{bmatrix} \\ &\equiv \boldsymbol{\alpha}^* \end{aligned} \qquad [12.A.5]$$

and

$$\begin{aligned} \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' &= \sigma^2\begin{bmatrix} 1/(\nu + T) & -\mathbf{1}'/(\nu + T) \\ \mathbf{1}/\nu & \mathbf{I}_T \end{bmatrix}\begin{bmatrix} \nu/(\nu + T) & \mathbf{1}' \\ -\mathbf{1}/(\nu + T) & \mathbf{I}_T \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2/(\nu + T) & \mathbf{0}' \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu) \end{bmatrix} \\ &\equiv \boldsymbol{\Sigma}^*. \end{aligned} \qquad [12.A.6]$$

Thus,

$$\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} = \boldsymbol{\alpha}'\mathbf{A}'(\mathbf{A}')^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{A}\boldsymbol{\alpha} = (\mathbf{A}\boldsymbol{\alpha})'(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')^{-1}(\mathbf{A}\boldsymbol{\alpha}) = \boldsymbol{\alpha}^{*\prime}(\boldsymbol{\Sigma}^*)^{-1}\boldsymbol{\alpha}^*. \qquad [12.A.7]$$

Moreover, observe that $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \begin{bmatrix} 1 & -\mathbf{1}'/(\nu + T) \\ \mathbf{0} & \mathbf{I}_T \end{bmatrix}\begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{1} & \mathbf{I}_T \end{bmatrix}.$$

Each of these triangular matrices has 1s along the principal diagonal and so has unit determinant, implying that $|\mathbf{A}| = 1$. Hence,

$$|\boldsymbol{\Sigma}^*| = |\mathbf{A}|\cdot|\boldsymbol{\Sigma}|\cdot|\mathbf{A}'| = |\boldsymbol{\Sigma}|. \qquad [12.A.8]$$

Substituting [12.A.5] through [12.A.8] into [12.A.1] gives

$$\begin{aligned} &f(\mathbf{y}, \mu; \sigma^2) \\ &= \frac{1}{(2\pi)^{(T+1)/2}} |\boldsymbol{\Sigma}^*|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}^{*\prime}(\boldsymbol{\Sigma}^*)^{-1}\boldsymbol{\alpha}^*\right\} \\ &= \frac{1}{(2\pi)^{(T+1)/2}} \left|\begin{matrix} \sigma^2/(\nu + T) & \mathbf{0}' \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu) \end{matrix}\right|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}\begin{bmatrix} \mu - m^* \\ \mathbf{y} - m\cdot\mathbf{1} \end{bmatrix}'\begin{bmatrix} \sigma^2/(\nu + T) & \mathbf{0}' \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu) \end{bmatrix}^{-1} \right. \\ &\quad \left. \times \begin{bmatrix} \mu - m^* \\ \mathbf{y} - m\cdot\mathbf{1} \end{bmatrix}\right\} \end{aligned} \qquad [12.A.9]$$

$$= \frac{1}{(2\pi)^{(T+1)/2}} \left[\frac{\sigma^2}{\nu + T}\right]^{-1/2} \cdot \left|\sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu)\right|^{-1/2}$$
$$\times \exp\left\{\frac{-(\mu - m^*)^2}{2\sigma^2/(\nu + T)} - \frac{(\mathbf{y} - m\cdot\mathbf{1})'(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu)^{-1}(\mathbf{y} - m\cdot\mathbf{1})}{2\sigma^2}\right\},$$

from which the factorization in Proposition 12.1 follows immediately. ■

■ **Proof of Proposition 12.2.** The product of [12.1.10] and [12.1.11] can be written as

$$f(\mathbf{y}, \boldsymbol{\beta}|\mathbf{X}; \sigma^2) = \frac{1}{(2\pi)^{(T+k)/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right\}$$

with

$$\underset{(T+k)\times 1}{\boldsymbol{\alpha}} \equiv \begin{bmatrix} \boldsymbol{\beta} - \mathbf{m} \\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \end{bmatrix}$$
$$\underset{(T+k)\times(T+k)}{\boldsymbol{\Sigma}} \equiv \begin{bmatrix} \sigma^2\mathbf{M} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}_T \end{bmatrix}.$$

As in the proof of Proposition 12.1, define

$$\underset{(T+k)\times(T+k)}{\mathbf{A}} = \begin{bmatrix} \mathbf{I}_k & -(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ \mathbf{0} & \mathbf{I}_T \end{bmatrix} \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{X} & \mathbf{I}_T \end{bmatrix}$$
$$= \begin{bmatrix} (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{M}^{-1} & -(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ \mathbf{X} & \mathbf{I}_T \end{bmatrix}.$$

Thus, $\mathbf{A}$ has unit determinant and

$$\mathbf{A}\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\beta} - \mathbf{m}^* \\ \mathbf{y} - \mathbf{X}\mathbf{m} \end{bmatrix}$$

with

$$\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \begin{bmatrix} \sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}') \end{bmatrix}.$$

Thus, as in equation [12.A.9],

$$f(\mathbf{y}, \boldsymbol{\beta}|\mathbf{X}; \sigma^2) = \frac{1}{(2\pi)^{(T+k)/2}} \left|\begin{matrix} \sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}') \end{matrix}\right|^{-1/2}$$
$$\times \exp\left\{-\frac{1}{2}\begin{bmatrix} \boldsymbol{\beta} - \mathbf{m}^* \\ \mathbf{y} - \mathbf{X}\mathbf{m} \end{bmatrix}' \begin{bmatrix} \sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}') \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\beta} - \mathbf{m}^* \\ \mathbf{y} - \mathbf{X}\mathbf{m} \end{bmatrix}\right\}. \quad ■$$

■ **Proof of Proposition 12.3(a).** We have that

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^{-2}|\mathbf{X}) = f(\mathbf{y}|\boldsymbol{\beta}, \sigma^{-2}, \mathbf{X}) \cdot f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{X}) \cdot f(\sigma^{-2}|\mathbf{X}). \qquad [12.A.10]$$

The first two terms on the right side are identical to [12.1.10] and [12.1.11]. Thus, Proposition 12.2 can be used to write [12.A.10] as

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^{-2}|\mathbf{X})$$
$$= \left\{\frac{1}{(2\pi\sigma^2)^{k/2}} |\mathbf{M}^*|^{1/2} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}\right\}$$
$$\times \left\{\frac{1}{(2\pi\sigma^2)^{T/2}} |\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'|^{-1/2}\right. \qquad [12.A.11]$$
$$\left.\times \exp\left\{[-1/(2\sigma^2)](\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m})\right\}\right\}$$
$$\times \left\{\frac{(\lambda/2)^{N/2}\sigma^{-2[(N/2)-1]} \exp[-\lambda\sigma^{-2}/2]}{\Gamma(N/2)}\right\}.$$

Define

$$\lambda^* \equiv \lambda + (\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m}); \qquad [12.A.12]$$

we will show later that this is the same as the value λ^* described in the proposition. For $N^* \equiv N + T$, the density [12.A.11] can be written as

$$\begin{aligned} &f(\mathbf{y}, \boldsymbol{\beta}, \sigma^{-2}|\mathbf{X}) \\ &= \left\{\frac{1}{(2\pi\sigma^2)^{k/2}}|\mathbf{M}^*|^{-1/2}\exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}\right\} \\ &\quad \times \left\{\frac{\sigma^{-2[(N^*/2)-1]}(\lambda/2)^{N/2}}{(2\pi)^{T/2}\Gamma(N/2)}|\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'|^{-1/2}\exp\left[-\frac{\lambda^*\sigma^{-2}}{2}\right]\right\} \qquad [12.A.13] \\ &= \left\{\frac{1}{(2\pi\sigma^2)^{k/2}}|\mathbf{M}^*|^{-1/2}\exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}\right\} \\ &\quad \times \left\{\frac{\sigma^{-2[(N^*/2)-1]}(\lambda^*/2)^{N^*/2}}{\Gamma(N^*/2)}\exp\left[-\frac{\lambda^*\sigma^{-2}}{2}\right]\right\} \\ &\quad \times \left\{\frac{\Gamma(N^*/2)(\lambda/2)^{N/2}}{(2\pi)^{T/2}\Gamma(N/2)(\lambda^*/2)^{N^*/2}}|\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'|^{-1/2}\right\}. \end{aligned}$$

The second term does not involve $\boldsymbol{\beta}$, and the third term does not involve $\boldsymbol{\beta}$ or σ^{-2}. Thus, [12.A.13] provides the factorization

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^{-2}|\mathbf{X}) = \{f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X})\}\cdot\{f(\sigma^{-2}|\mathbf{y}, \mathbf{X})\}\cdot\{f(\mathbf{y}|\mathbf{X})\},$$

where $f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X})$ is a $N(\mathbf{m}^*, \sigma^2\mathbf{M}^*)$ density, $f(\sigma^{-2}|\mathbf{y}, \mathbf{X})$ is a $\Gamma(N^*, \lambda^*)$ density, and $f(\mathbf{y}|\mathbf{X})$ can be written as

$$\begin{aligned} f(\mathbf{y}|\mathbf{X}) &= \left\{\frac{\Gamma(N^*/2)(\lambda/2)^{N/2}}{(2\pi)^{T/2}\Gamma(N/2)(\lambda^*/2)^{N^*/2}}|\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'|^{-1/2}\right\} \\ &= \left\{\frac{\Gamma[(N+T)/2]\lambda^{N/2}|\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'|^{-1/2}}{\pi^{T/2}\Gamma(N/2)\{\lambda + (\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m})\}^{(N+T)/2}}\right\} \\ &= c\cdot\{1 + (1/N)(\mathbf{y} - \mathbf{X}\mathbf{m})'[(\lambda/N)(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')]^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m})\}^{-(N+T)/2}, \end{aligned}$$

where

$$c = \frac{\Gamma[(N+T)/2](1/N)^{T/2}|(\lambda/N)(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')|^{-1/2}}{\pi^{T/2}\Gamma(N/2)}.$$

Thus, $f(\mathbf{y}|\mathbf{X})$ is a T-dimensional Student's t density with N degrees of freedom, mean $\mathbf{X}\mathbf{m}$, and scale matrix $(\lambda/N)(\mathbf{I}_T + \mathbf{X}\mathbf{M}'\mathbf{X}')$. Hence, the distributions of $(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X})$ and $(\sigma^{-2}|\mathbf{y}, \mathbf{X})$ are as claimed in Proposition 12.3, provided that the magnitude λ^* defined in [12.A.12] is the same as the expression in [12.1.28]. To verify that this is indeed the case, notice that

$$(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')^{-1} = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}', \qquad [12.A.14]$$

as can be verified by premultiplying [12.A.14] by $(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')$:

$$\begin{aligned} &(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'] \\ &\quad = \mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}' - \mathbf{X}\mathbf{M}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}' \\ &\quad = \mathbf{I}_T + \mathbf{X}\left\{\mathbf{M}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1}) - \mathbf{I}_k - \mathbf{M}(\mathbf{X}'\mathbf{X})\right\}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}' \\ &\quad = \mathbf{I}_T. \end{aligned}$$

Using [12.A.14], we see that

$$\begin{aligned} &(\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m}) \\ &\quad = (\mathbf{y} - \mathbf{X}\mathbf{m})'[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'](\mathbf{y} - \mathbf{X}\mathbf{m}) \\ &\quad = (\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{m})'[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'](\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{m}) \\ &\quad = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \mathbf{m})'\mathbf{X}'[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}']\mathbf{X}(\mathbf{b} - \mathbf{m}). \end{aligned} \qquad [12.A.15]$$

where cross-product terms have disappeared because of the *OLS* orthogonality condition $(\mathbf{y} - \mathbf{Xb})'\mathbf{X} = \mathbf{0}'$. Furthermore,

$$\begin{aligned}
&\mathbf{X}'[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}']\mathbf{X} \\
&\quad = [\mathbf{I}_k - (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}]\mathbf{X}'\mathbf{X} \\
&\quad = [(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1} - (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}]\mathbf{X}'\mathbf{X} \\
&\quad = \mathbf{M}^{-1}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'\mathbf{X}.
\end{aligned}$$

This allows [12.A.15] to be written as

$$\begin{aligned}
(\mathbf{y} - \mathbf{Xm})'(\mathbf{I}_T + \mathbf{XMX}')^{-1}(\mathbf{y} - \mathbf{Xm}) \\
= (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) + (\mathbf{b} - \mathbf{m})'\mathbf{M}^{-1}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{m}),
\end{aligned}$$

establishing the equivalence of [12.A.12] and [12.1.28].

Proof of (b). The joint posterior density of $\boldsymbol{\beta}$ and σ^{-2} is given by

$$\begin{aligned}
&f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{y}, \mathbf{X}) \\
&\quad = f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X}) \cdot f(\sigma^{-2}|\mathbf{y}, \mathbf{X}) \\
&\quad = \left\{\frac{1}{(2\pi\sigma^2)^{k/2}}|\mathbf{M}^*|^{-1/2}\exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}\right\} \\
&\qquad \times \left\{\frac{\sigma^{-2[(N^*/2)-1]}(\lambda^*/2)^{N^*/2}}{\Gamma(N^*/2)}\exp[-\lambda^*\sigma^{-2}/2]\right\} \\
&\quad = \left(\frac{\sigma^{-2[(k+N^*)/2]-1]}}{\Gamma[(k + N^*)/2]} \times \left\{\frac{\lambda^*}{2}\cdot[1 + (\boldsymbol{\beta} - \mathbf{m}^*)'(\lambda^*\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]\right\}^{(k+N^*)/2}\right. \\
&\qquad \left. \times \exp\left\{-\frac{\lambda^*}{2}\cdot[1 + (\boldsymbol{\beta} - \mathbf{m}^*)'(\lambda^*\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]\sigma^{-2}\right\}\right) \\
&\qquad \times \left\{\frac{\Gamma[(k + N^*)/2]}{(\lambda^*)^{k/2}\pi^{k/2}\Gamma(N^*/2)}|\mathbf{M}^*|^{-1/2}[1 + (\boldsymbol{\beta} - \mathbf{m}^*)'(\lambda^*\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]^{-(k+N^*)/2}\right\} \\
&\quad = \{f(\sigma^{-2}|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})\}\cdot\{f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})\},
\end{aligned}$$

where $f(\sigma^{-2}|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ will be recognized as a $\Gamma\big((k + N^*), \lambda^*[1 + (\boldsymbol{\beta} - \mathbf{m}^*)'(\lambda^*\mathbf{M}^*)^{-1} \times (\boldsymbol{\beta} - \mathbf{m}^*)]\big)$ density, while $f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ can be written as

$$\begin{aligned}
f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \left\{\frac{\Gamma[(k + N^*)/2]}{(N^*)^{k/2}\pi^{k/2}\Gamma(N^*/2)}|(\lambda^*/N^*)\mathbf{M}^*|^{-1/2}\right. \\
\left. \times [1 + (1/N^*)(\boldsymbol{\beta} - \mathbf{m}^*)'[(\lambda^*/N^*)\mathbf{M}^*]^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]^{-(k+N^*)/2}\right\},
\end{aligned}$$

which is a k-dimensional t density with N^* degrees of freedom, mean $\mathbf{m}^*$, and scale matrix $(\lambda^*/N^*)\mathbf{M}^*$.

Proof of (c). Notice that conditional on $\mathbf{y}$, $\mathbf{X}$, and σ^2, the variable

$$Z = [\mathbf{R}(\boldsymbol{\beta} - \mathbf{m}^*)]'[\sigma^2\mathbf{R}(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\cdot[\mathbf{R}(\boldsymbol{\beta} - \mathbf{m}^*)]$$

is distributed $\chi^2(m)$, from Proposition 8.1. The variable Q in [12.1.30] is equal to $Z\cdot\sigma^2N^*/(m\lambda^*)$, and so conditional on $\mathbf{y}$, $\mathbf{X}$, and σ^2, the variable Q is distributed $\Gamma\big(m, (m\lambda^*)/(\sigma^2N^*)\big)$:

$$f(q|\sigma^{-2}, \mathbf{y}, \mathbf{X}) = \frac{[m\lambda^*/(2\sigma^2N^*)]^{m/2}q^{[(m/2)-1]}\exp[-m\lambda^*q/(2\sigma^2N^*)]}{\Gamma(m/2)}. \quad [12.A.16]$$

The joint posterior density of q and σ^{-2} is

$$
\begin{aligned}
f(q, \sigma^{-2}|\mathbf{y}, \mathbf{X}) &= f(q|\sigma^{-2}, \mathbf{y}, \mathbf{X}) \cdot f(\sigma^{-2}|\mathbf{y}, \mathbf{X}) \\
&= \left\{ \frac{[m\lambda^*/(2\sigma^2 N^*)]^{m/2} q^{[(m/2)-1]} \exp[-m\lambda^* q/(2\sigma^2 N^*)]}{\Gamma(m/2)} \right\} \\
&\quad \times \left\{ \frac{\sigma^{-2[(N^*/2)-1]} (\lambda^*/2)^{N^*/2}}{\Gamma(N^*/2)} \exp[-\lambda^* \sigma^{-2}/2] \right\} \\
&= \left\{ \frac{[(N^* + mq) \cdot [\lambda^*/(2N^*)]]^{[(N^*+m)/2]}}{\Gamma[(N^* + m)/2]} \right. \qquad [12.A.17] \\
&\quad \left. \times \sigma^{-2[(m+N^*)/2]-1} \exp[-(N^* + mq)(\lambda^*/N^*)\sigma^{-2}/2] \right\} \\
&\quad \times \left\{ \frac{m^{m/2}(N^*)^{N^*/2}\Gamma[(N^* + m)/2] q^{[(m/2)-1]}}{\Gamma(m/2)\Gamma(N^*/2)(N^* + mq)^{[(N^*+m)/2]}} \right\} \\
&= \{f(\sigma^{-2}|q, \mathbf{y}, \mathbf{X})\} \cdot \{f(q|\mathbf{y}, \mathbf{X})\},
\end{aligned}
$$

where $f(\sigma^{-2}|q, \mathbf{y}, \mathbf{X})$ is a $\Gamma\big((m + N^*), (N^* + mq)(\lambda^*/N^*)\big)$ density and $f(q|\mathbf{y}, \mathbf{X})$ is an $F(m, N^*)$ density. ■

Chapter 12 Exercise

12.1. Deduce Proposition 12.1 as a special case of Proposition 12.2.

Chapter 12 References

Almon, Shirley. 1965. "The Distributed Lag between Capital Appropriations and Expenditures." *Econometrica* 33:178–96.

Barro, Robert J. 1979. "On the Determination of the Public Debt." *Journal of Political Economy* 87:940–71.

DeGroot, Morris H. 1970. *Optimal Statistical Decisions*. New York: McGraw-Hill.

Diebold, Francis X., and James A. Nason. 1990. "Nonparametric Exchange Rate Prediction?" *Journal of International Economics* 28:315–32.

Doan, Thomas A. 1990. *RATS User's Manual*. VAR Econometrics, Suite 612, 1800 Sherman Ave., Evanston, IL 60201.

Fama, Eugene F. 1965. "The Behavior of Stock Market Prices." *Journal of Business* 38:34–105.

Geweke, John. 1988a. "The Secular and Cyclical Behavior of Real GDP in 19 OECD Countries, 1957–1983." *Journal of Business and Economic Statistics* 6:479–86.

———. 1988b. "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference." *Journal of Econometrics* 38:73–89.

———. 1989. "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica* 57:1317–39.

Hall, Robert E. 1978. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *Journal of Political Economy* 86:971–87.

Hammersley, J. M., and D. C. Handscomb. 1964. *Monte Carlo Methods*, 1st ed. London: Methuen.

Hoerl, A. E., and R. W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12:55–82.

Kloek, T., and H. K. van Dijk. 1978. "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo." *Econometrica* 46:1–19.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Litterman, Robert B. 1986. "Forecasting with Bayesian Vector Autoregressions—Five Years of Experience." *Journal of Business and Economic Statistics* 4:25–38.

Mankiw, N. Gregory. 1987. "The Optimal Collection of Seigniorage: Theory and Evidence." *Journal of Monetary Economics* 20:327–41.

Rothenberg, Thomas J. 1973. *Efficient Estimation with A Priori Information*. New Haven, Conn.: Yale University Press.

Shiller, Robert J. 1973. "A Distributed Lag Estimator Derived from Smoothness Priors." *Econometrica* 41:775–88.

Theil, Henri. 1971. *Principles of Econometrics*. New York: Wiley.

Tierney, Luke, and Joseph B. Kadane. 1986. "Accurate Approximations for Posterior Moments and Marginal Densities." *Journal of the American Statistical Association* 81:82–86.

Zellner, Arnold. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.