Some Background on Bayesian Statistics and Econometrics

Economics 721
John C. Chao

I. Some Rudimentary Bayesian Statistics

- Bayes' Rule

  Bayesian statistics is so named because, under this approach, statistical inference is based on the *posterior distribution* obtained via Bayes' rule

$$p(\theta|\,x) = \frac{f(x|\theta)\pi(\theta)}{m(x)},$$

where

$f(x|\theta)$ - data density,

$\pi(\theta)$ - prior density,

$m(x)$ - marginal density of the data,

$p(\theta|\,x)$- posterior density,

$\theta \in \Theta$, where Θ denotes the parameter space.

- **Remarks**:
  - **(i)** $m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta.$
  - **(ii)** Typically, in Bayesian statistics, we focus on a parametric framework so that Θ is finite dimensional.
  - **(iii)** Note that $f(x|\theta) = l(\theta, x)$, so that $f(x|\theta)$ is just the likelihood function when reinterpreted as a function of θ given x.
  - **(iv)** If $\pi(\theta) = c$, for some constant c, then

$$p(\theta|\, x) \propto l(\theta, x).$$

- **Some special features of Bayesian inference**:
  - **(i)** Parameter θ is random.
  - **(ii)** Inference is made *conditional* on the data.
  - **(iii)** Model specification requires both specification of the likelihood and of the prior.

- Point Estimation

a. Point estimate of θ can be obtained by taking the mean, median, or mode of the posterior distribution.

b. More generally, given a loss function $L\left(\theta, \widehat{\theta}\right)$, point estimates of θ can be obtained by miminizing the expected loss à posteriori, i.e.,

$$\begin{aligned}\widehat{\theta} &= \arg\min E^{\pi}\left[L\left(\theta, \widehat{\theta}\right) | x\right] \\ &= \arg\min \int_{\Theta} L\left(\theta, \widehat{\theta}\right) p(\theta | x) d\theta\end{aligned}$$

c. Some common loss functions:

1. **quadratic loss**:

$$L\left(\theta, \widehat{\theta}\right) = \left(\theta - \widehat{\theta}\right)^2$$

(**Note**: Use of quadratic loss results in $\widehat{\theta}$ = posterior mean.)

2. **absolute error loss**:

$$L\left(\theta, \widehat{\theta}\right) = \left|\theta - \widehat{\theta}\right|$$

(**Note**: Use of absolute erro loss results in $\widehat{\theta}$ = posterior median.)

- Interval (or Set) Estimation

a. **Bayesian Credible Set**: A $100(1-\alpha)\%$ credible set for θ is a subset C of Θ such that

$$1-\alpha \le p(C|x) = \int_C p(\theta|x)d\theta$$

(**Note**: A problem with the above definition is that the set C is in most cases not unique.)

b. **Highest Posterior Density (HPD) Credible Set**: The $100(1-\alpha)\%$ HPD credible set for θ is the subset C of Θ of the form

$$C = \{\theta \in \Theta : p(\theta|x) \ge k(\alpha)\},$$

where $k(\alpha)$ is the largest constant such that

$$p(C|x) \ge 1-\alpha$$

- Hypothesis Testing

a. **Posterior Odds Ratio and Bayes Factor**:

Consider the testing problem

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 forms a partition of the parameter space Θ (i.e., $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \phi$). A Bayes test of these hypotheses is based on the posterior odds ratio

$$PO = \frac{p(H_0|x)}{p(H_1|x)} = \frac{p(\theta \in \Theta_0|x)}{p(\theta \in \Theta_1|x)}$$

or the Bayes factor

$$B = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}} = \frac{p(H_0|x)/p(H_1|x)}{\pi(H_0)/\pi(H_1)} = \frac{p(H_0|x)\pi(H_1)}{p(H_1|x)\pi(H_0)}$$

b. More explicitly, we often write the prior as

$$\pi(\theta) = \begin{cases} \pi_0 g_0(\theta) & \text{if } \theta \in \Theta_0 \\ \pi_1 g_1(\theta) & \text{if } \theta \in \Theta_1 \end{cases},$$

where g_0 and g_1 are (proper) densities which describe how the prior mass is spread out over the two hypotheses and where $\pi_0 = \pi(H_0)$ and $\pi_1 = \pi(H_1)$ are the prior probabilities of H_0 and H_1. Hence,

$$PO = \frac{\int_{\Theta_0} f(x|\theta)\pi_0 g_0(\theta)d\theta/m(x)}{\int_{\Theta_1} f(x|\theta)\pi_1 g_1(\theta)d\theta/m(x)}$$

$$= \frac{\pi_0 \int_{\Theta_0} f(x|\theta)g_0(\theta)d\theta}{\pi_1 \int_{\Theta_1} f(x|\theta)g_1(\theta)d\theta} = \frac{ML_0}{ML_1}$$

and

$$B = \frac{\int_{\Theta_0} f(x|\theta) g_0(\theta) d\theta}{\int_{\Theta_1} f(x|\theta) g_1(\theta) d\theta}.$$

c. Remarks:

(i) Note that the Bayes factor is some sense a "weighted likelihood ratio".

(ii) Typically, we take $\pi_0 = \pi_1 = \frac{1}{2}$, so $PO = B$.

(iii) Note that one should avoid specifying $g_0(\theta)$ and $g_1(\theta)$ as improper priors. Too see why, suppose that the parameter spaces Θ_0 and Θ_1 are unbounded and let $g_0(\theta) = c_0$ and $g_1(\theta) = c_1$; then,

$$B = \frac{\int_{\Theta_0} f(x|\theta) g_0(\theta) d\theta}{\int_{\Theta_1} f(x|\theta) g_1(\theta) d\theta}$$

$$= \frac{c_0 \int_{\Theta_0} f(x|\theta) d\theta}{c_1 \int_{\Theta_1} f(x|\theta) d\theta}.$$

Hence, we can make B as big or small as we wish by manipulating the height of the densities c_0 and c_1, since there is no constant of normalization for improper priors.

d. **Decision Rule**:

Consider the case where $\pi_0 = \pi_1 = \frac{1}{2}$, then the decision rule is

$$\text{Reject } H_0 \text{ if } PO < 1,$$

$$\text{Accept } H_0 \text{ if } PO \geq 1.$$

e. **Further Remark**: Note the symmetric way in which H_0 and H_1 are treated in contrast to frequentist significant testing.

- Bayesian Handling of Nuisance Parameters:

  a. Marginalization - i.e., nuisance parameters are integrated out.

  b. To illustrate, suppose that we can partition

$$\theta = (\theta_1', \theta_2')',$$

  where

  θ_1 - parameter (vector) of interest,

  θ_2 - nuisance parameter (vector).

  Bayesian inference is based on the marginal posterior distribution of θ_1 obtained by integrating out θ_2, i.e.,

$$p(\theta_1 | x) = \int_{\Theta_2} p(\theta_1, \theta_2 | x) d\theta_2.$$

II. An Illustrative Example - Linear Regression Model

Consider the linear model

$$\underset{T\times 1}{y} = \underset{T\times k}{X}\,\underset{k\times 1}{\beta} + \underset{T\times 1}{u},\; u \sim N(0,\sigma^2 I_T).$$

A. Diffuse Prior Analysis:

1. Diffuse Prior

$$\pi(\beta,\sigma) \propto \frac{1}{\sigma} \text{ for } 0 < \sigma < \infty$$

Remarks:

(i) This prior is uniform on β.

(ii) This prior is also uniform on $\theta = \ln\sigma$, so that the prior density takes the form

$$\pi(\theta) \propto 1.$$

It follows that

$$\pi(\sigma) \propto \left|\frac{d\theta}{d\sigma}\right| = \frac{1}{\sigma}.$$

2. Likelihood Function

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

3. Joint Posterior Distribution

$$p(\beta,\sigma|y,X)$$

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

$$= \frac{1}{(2\pi)^{\frac{T}{2}}\sigma^{T+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[vs^2 + \left(\beta - \widehat{\beta}\right)'X'X\left(\beta - \widehat{\beta}\right)\right]\right\},$$

where

$$\widehat{\beta} = (X'X)^{-1}X'y,$$

$$s^2 = \frac{\left(y - X\widehat{\beta}\right)'\left(y - X\widehat{\beta}\right)}{\nu},$$

$$\nu = T - k.$$

4. Marginal Posterior of β

$$p(\beta|y,X) = \int_0^\infty p(\beta,\sigma|y,X)d\sigma$$

$$\propto \left|\nu s^2 + \left(\beta - \widehat{\beta}\right)' X'X\left(\beta - \widehat{\beta}\right)\right|^{-\frac{T}{2}}$$

Note: this marginal posterior of β has a multivariate t distribution with $T - k$ degrees of freedom.

5. **Remark**: Note that

$$p(\beta|y,X) \rightarrow \text{ normal density as } T \rightarrow \infty.$$

This result does not depend on whether X consider on lagged dependent variable or not. In particular, consider the special case of a $AR(1)$ model, i.e.,

$$y_t = \beta y_{t-1} + u_t,\ i.i.d.N(0,\sigma^2).$$

The asymptotic normality of the marginal posterior of β holds even if $\beta_0 = 1$, i.e., even if we have a unit root model. THIS IS VERY DIFFERENT FROM RESULTS OBTAINED UNDER THE FREQUENTIST OR CLASSICAL APPROACH.

B. Gaussian Prior Analysis

1. Prior Specification

$$\pi(\beta|\sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}\sigma^k} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \overline{\beta})' V_\beta (\beta - \overline{\beta})\right\},$$

$$\pi(\sigma) = \frac{1}{\sigma}.$$

2. Marginal Posterior of β

$$p(\beta|y, X) \propto \left| T\tilde{s}^2 + \left(\beta - \tilde{\beta}\right)' [X'X + V_\beta] \left(\beta - \tilde{\beta}\right) \right|^{-\frac{T+k}{2}},$$

where

$$\tilde{\beta} = [X'X + V_\beta]^{-1}[X'y + V_\beta\overline{\beta}],$$

$$\tilde{s}^2 = \frac{1}{T}\left[y'y + \overline{\beta}' V_\beta \overline{\beta} - \tilde{\beta}'(X'X + V_\beta)\tilde{\beta}\right]$$

3. **Remark**: Note that we can rewrite

$$\widetilde{\beta} = [X'X + V_\beta]^{-1}(X'X)\widehat{\beta} + [X'X + V_\beta]^{-1}V_\beta\overline{\beta},$$

where $\widehat{\beta} = (X'X)^{-1}X'y$. In the case where $k = 1$,

$$\widetilde{\beta} = \left(\frac{x'x}{x'x + v_\beta}\right)\widehat{\beta} + \left(\frac{v_\beta}{x'x + v_\beta}\right)\overline{\beta},$$

so $\widetilde{\beta}$ is a linear combination of the *MLE* estimator $\widehat{\beta}$ and the prior mean $\overline{\beta}$.

C. Model Selection and Hypothesis Testing

Consider the problem of selecting the lag order of an autoregression

$$M_k : y_t = \beta_1 y_{t-1} + \ldots + \beta_k y_{t-k} + u_t,$$

where $\{u_t\} \equiv i.i.d. N(0, \sigma^2)$.

1. Prior Specification Based on Training Sample:

$$\pi(\beta|\sigma, M_k)$$

$$\propto \frac{1}{\sigma^k} \exp\left\{-\frac{1}{2\sigma^2} e(k)' X_1(k)' X_1(k) e(k)\right\}$$

$$\pi(\sigma|M_k)$$

$$\propto \frac{1}{\sigma^{T_1-k+1}} \exp\left\{-\frac{(T_1-k)s_1^2(k)}{2\sigma^2}\right\},$$

where

$$e(k) = \beta(k) - \widehat{\beta}_1(k),$$

$$s_1^2(k) = \frac{\widehat{u}_1(k)'\widehat{u}_1(k)}{T_1 - k},$$

$$
\begin{aligned}
Y_1 &= (y_1, \ldots, y_{T_1})', \\
X_1(k) &= (x_1(k), \ldots., x_{T_1}(k))', \\
x_t(k) &= (y_{t-1}, \ldots.., y_{t-k})', \\
\beta(k) &= (\beta_1, \ldots., \beta_k)', \\
\widehat{\beta}_1(k) &= [X_1(k)'X_1(k)]^{-1}X_1(k)'Y_1, \\
\widehat{u}_1(k) &= Y_1 - X_1(k)\widehat{\beta}_1(k).
\end{aligned}
$$

2. Likelihood Function

$$
\begin{aligned}
& f(Y_2|\beta, \sigma, X_2, M_k) \\
& \propto \frac{1}{\sigma^{T_2}} \exp\left\{-\frac{1}{2\sigma^2} u_2(k)' u_2(k)\right\},
\end{aligned}
$$

where

$$
\begin{aligned}
u_2(k) &= Y_2 - X_2(k)\beta(k), \\
Y_2 &= (y_{T_1+1}, \ldots, y_T)', \\
X_2(k) &= (x_{T_1+1}(k), \ldots., x_T(k))'.
\end{aligned}
$$

3. Marginal Likelihood

$$
\begin{aligned}
Cr(k) &= -\frac{2}{T}\ln(ML_k) \\
&= \ln(s^2(k)) + \frac{1}{T}\ln|X(k)'X(k)| \\
&\quad + \text{lower order terms},
\end{aligned}
$$

where

$$
\begin{aligned}
s^2(k) &= \frac{\hat{u}(k)'\hat{u}(k)}{T-k}, \\
\hat{u}(k) &= Y - X(k)\hat{\beta}(k), \\
Y &= (y_1, \ldots, y_T)', \\
X(k) &= (x_1(k), \ldots., x_T(k))', \\
\hat{\beta}(k) &= [X(k)'X(k)]^{-1}X(k)'Y.
\end{aligned}
$$

Further Approximation:

$$
\begin{aligned}
Cr(k) &= \ln(s^2(k)) + \frac{1}{T}\ln|X(k)'X(k)| \\
&\quad + \text{lower order terms} \\
&= \ln(s^2(k)) + \frac{1}{T}\ln T^k|X(k)'X(k)/T| \\
&\quad + \text{lower order terms} \\
&= \ln(s^2(k)) + \frac{k}{T}\ln T \\
&\quad + \frac{1}{T}\ln|X(k)'X(k)/T| \\
&\quad + \text{lower order terms} \\
&= BIC(k) + \text{lower order terms.}
\end{aligned}
$$

III. Some Aspects of Bayesian Computation

- Laplace's Method
  Consider the integral

$$\int b(\theta)\exp\{-nh(\theta)\}d\theta,$$

where $h(\theta) = n^{-1}\ln(l(\theta)\pi(\theta))$. Laplace's method approximates the integral above by expanding the integrand as a Taylor series and then integrate with respect to the quadratic term, i.e.,

$$\int \left(b(\hat{\theta}) + b'(\hat{\theta})\left(\theta - \hat{\theta}\right) + \ldots \right)$$

$$\exp\left\{ -n\left[h\left(\hat{\theta}\right) + \frac{1}{2} h''\left(\hat{\theta}\right)\left(\theta - \hat{\theta}\right)^2 \right.\right.$$

$$\left.\left. + \ldots \right]\right\} d\theta$$

$$= \exp\left\{ -nh\left(\hat{\theta}\right) \right\} (2\pi)^{\frac{1}{2}} \left[nh''\left(\hat{\theta}\right) \right]^{-\frac{1}{2}}$$

$$\int \left(b(\hat{\theta}) + b'(\hat{\theta})\left(\theta - \hat{\theta}\right) + \ldots \right)$$

$$(2\pi)^{-\frac{1}{2}} \sqrt{nh''\left(\hat{\theta}\right)} \exp\left\{ -\frac{1}{2} nh''\left(\hat{\theta}\right)\left(\theta - \hat{\theta}\right)^2 \right\}$$

$$(1 + \text{lower order terms}) d\theta$$

$$= \exp\left\{ -nh\left(\hat{\theta}\right) \right\} (2\pi)^{\frac{1}{2}} \left[nh''\left(\hat{\theta}\right) \right]^{-\frac{1}{2}} b(\hat{\theta})$$

$$(1 + \text{lower order terms}),$$

where $\hat{\theta}$ denotes the maximum of $-h$.

- **Example**: (Chao and Phillips, 1999) Consider joint estimation of cointegrating rank and lag order in the vector error-correction model

$$\begin{aligned}\Delta y_t &= \gamma\beta' y_{t-1} + \Phi_1 \Delta y_{t-1} + .. + \Phi_p \Delta y_{t-p} + \varepsilon_t \\ &= G u_t + \varepsilon_t,\end{aligned}$$

where

$$\begin{aligned}u_t &= \begin{bmatrix} \beta' y_{t-1} \\ z_t \end{bmatrix}, \\ G &= [\gamma, \Phi], \\ \Phi &= [\Phi_1, \ldots, \Phi_p] \\ z_t &= [\Delta y'_{t-1}, \ldots, \Delta y_{t-p}]'.\end{aligned}$$

Laplace method was used to construct the PIC criterion which has interpretation as a transformed marginal likelihood

$$PIC(p,r) = \ln\left|\widehat{\Sigma}\right| + \frac{1}{n}\ln\left|\widehat{B}_n\right|.$$

- Monte Carlo Integration
  Consider the integral

$$E(\theta) = \int g(\theta)f(\theta)d\theta,$$

where $f(\theta)$ is a p.d.f. We can approximate for $E(\theta)$ by drawing $i.i.d.$ sample $\theta_1,\ldots,\theta_m$ from $f(\theta)$ and estimate $E(\theta)$ by

$$\widehat{E}(\theta) = \frac{1}{m}\sum_{i=1}^{m} g(\theta_i)$$

By the (strong) law of large number

$$\widehat{E}(\theta) \stackrel{a.s.}{\rightarrow} E(\theta)$$

Note that also that the convergence above does not depend on the dimension of θ.

- Importance Sampling

  The algorithm above may be inefficient because we may draw a lot of θ_i where $g(\theta)$ is close to zero. Alternatively, we can instead draw an $i.i.d.$ sample $\theta_1, \ldots, \theta_m$ from an importance function $I(\theta)$ which better mimic the function $g(\theta)$ and estimate $E(\theta)$ by

$$\widetilde{E}(\theta) = \frac{\sum_{i=1}^{m} g(\theta_i) w_i}{\sum_{i=1}^{m} w_i}$$

  where

$$w_i = \frac{f(\theta_i)}{I(\theta_i)}$$