Econometrics I

Professor William Greene
Stern School of Business
Department of Economics
Econometrics I

Part 24 – Bayesian Estimation
Bayesian Estimators

- “Random Parameters” vs. Randomly Distributed Parameters
- Models of Individual Heterogeneity
  - Random Effects: Consumer Brand Choice
  - Fixed Effects: Hospital Costs
Bayesian Estimation

- Specification of conditional likelihood: \( f(\text{data} \mid \text{parameters}) \)
- Specification of priors: \( g(\text{parameters}) \)
- Posterior density of parameters:
  \[
  f(\text{parameters} \mid \text{data}) = \frac{f(\text{data} \mid \text{parameters})g(\text{parameters})}{f(\text{data})}
  \]
- Posterior mean = \( E[\text{parameters} \mid \text{data}] \)
The Marginal Density for the Data is Irrelevant

\[ f(\beta | \text{data}) = \frac{f(\text{data} | \beta)p(\beta)}{f(\text{data})} = \frac{L(\text{data} | \beta)p(\beta)}{f(\text{data})} \]

Joint density of \( \beta \) and data is \( f(\text{data}, \beta) = L(\text{data} | \beta)p(\beta) \)

Marginal density of the data is

\[ f(\text{data}) = \int_{\beta} f(\text{data}, \beta)d\beta = \int_{\beta} L(\text{data} | \beta)p(\beta)d\beta \]

Thus, \( f(\beta | \text{data}) = \frac{L(\text{data} | \beta)p(\beta)}{\int_{\beta} L(\text{data} | \beta)p(\beta)d\beta} \)

Posterior Mean = \( \int_{\beta} p(\beta | \text{data})d\beta = \frac{\int_{\beta} \beta L(\text{data} | \beta)p(\beta)d\beta}{\int_{\beta} L(\text{data} | \beta)p(\beta)d\beta} \)

Requires specification of the likelihood and the prior.
Computing Bayesian Estimators

- First generation: Do the integration (math)

\[ E(\beta \mid \text{data}) = \int_\beta \beta \frac{f(\text{data} \mid \beta) g(\beta)}{f(\text{data})} d\beta \]

- Contemporary - Simulation:
  1. Deduce the posterior
  2. Draw random samples of draws from the posterior and compute the sample means and variances of the samples. (Relies on the law of large numbers.)
Modeling Issues

- As \( n \to \infty \), the likelihood dominates and the prior disappears \( \Rightarrow \) Bayesian and Classical MLE converge. (Needs the mode of the posterior to converge to the mean.)

- Priors
  - Diffuse \( \Rightarrow \) large variances imply little prior information. (NONINFORMATIVE)
A Practical Problem

Sampling from the joint posterior may be impossible. E.g., linear regression.

\[
f(\beta, \sigma^2 | y, X) \propto \frac{[\nu s^2]^{v+2}}{\Gamma(v + 2)} \left[ \frac{1}{\sigma^2} \right]^{v+1} e^{-\nu s^2(1/\sigma^2)} [2\pi]^{-K/2} | \sigma^2 (X'X)^{-1} |^{-1/2} \times \exp(-\frac{1}{2}(\beta - b)'[\sigma^2 (X'X)^{-1}]^{-1}(\beta - b))
\]

What is this???

To do 'simulation based estimation' here, we need joint observations on \((\beta, \sigma^2)\).
A Solution to the Sampling Problem

The joint posterior, \( p(\boldsymbol{\beta}, \sigma^2 | \text{data}) \) is intractable. But, for inference about \( \boldsymbol{\beta} \), a sample from the marginal posterior, \( p(\boldsymbol{\beta} | \text{data}) \) would suffice. For inference about \( \sigma^2 \), a sample from the marginal posterior of \( \sigma^2 \), \( p(\sigma^2 | \text{data}) \) would suffice.

Can we deduce these? For this problem, we do have conditionals:

\[
p(\boldsymbol{\beta} | \sigma^2, \text{data}) = N[\boldsymbol{b}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}] \\
p(\sigma^2 | \boldsymbol{\beta}, \text{data}) = K \times \frac{\sum_i (y_i - x_i\boldsymbol{\beta})^2}{\sigma^2} = \text{a gamma distribution}
\]

Can we use this information to sample from \( p(\boldsymbol{\beta} | \text{data}) \) and \( p(\sigma^2 | \text{data}) \)?
The Gibbs Sampler

- **Target:** Sample from marginals of $f(x_1, x_2) = \text{joint distribution}$
- Joint distribution is unknown or it is not possible to sample from the joint distribution.
- Assumed: $f(x_1|\cdot x_2)$ and $f(x_2|\cdot x_1)$ both known and samples can be drawn from both.
- **Gibbs sampling:** Obtain one draw from $x_1, x_2$ by many cycles between $x_1|x_2$ and $x_2|x_1$.
  - Start $x_{1,0}$ anywhere in the right range.
  - Draw $x_{2,0}$ from $x_2|x_{1,0}$.
  - Return to $x_{1,1}$ from $x_1|x_{2,0}$ and so on.
  - Several thousand cycles produces the draws
  - Discard the first several thousand to avoid initial conditions. (Burn in)
- **Average the draws to estimate the marginal means.**
Bivariate Normal Sampling

Draw a random sample from bivariate normal \[
\begin{pmatrix}
0 \\
0 \\
\rho
\end{pmatrix},
\begin{pmatrix}
1 & \rho \\
\rho & 1
\end{pmatrix}
\]

(1) Direct approach: \[
\begin{pmatrix}
v_1 \\
v_2
\end{pmatrix}_r = \Gamma \begin{pmatrix}
u_1 \\
u_2
\end{pmatrix}_r \quad \text{where} \begin{pmatrix}
u_1 \\
u_2
\end{pmatrix}
\]
are two independent standard normal draws (easy) and \[
\Gamma = \begin{pmatrix}
1 & 0 \\
\theta_1 & \theta_2
\end{pmatrix}
\]
such that \[\Gamma \Gamma' = \begin{pmatrix}
1 & \rho \\
\rho & 1
\end{pmatrix} \cdot \begin{pmatrix}
1 & 0 \\
\theta_1 & \theta_2
\end{pmatrix} = \theta_1 = \rho, \quad \theta_2 = \sqrt{1 - \rho^2}.
\]

(2) Gibbs sampler: \[v_1 \mid v_2 \sim N\left[\rho v_2, \sqrt{1 - \rho^2}\right]\]
\[v_2 \mid v_1 \sim N\left[\rho v_1, \sqrt{1 - \rho^2}\right]\]
Gibbs Sampling for the Linear Regression Model

\[ p(\beta | \sigma^2, \text{data}) = N[b, \sigma^2 (X'X)^{-1}] \]

\[ p(\sigma^2 | \beta, \text{data}) = K \times \frac{\sum_i (y_i - x_i'\beta)^2}{\sigma^2} \]

\[ = \text{a gamma distribution} \]

Iterate back and forth between these two distributions
Application – the Probit Model

(a) \( y_{i}^{*} = \mathbf{x}_{i}^{\prime} \mathbf{\beta} + \epsilon_{i} \quad \epsilon_{i} \sim \text{N}[0,1] \)

(b) \( y_{i} = 1 \text{ if } y_{i}^{*} > 0, \ 0 \text{ otherwise} \)

Consider estimation of \( \mathbf{\beta} \) and \( y_{i}^{*} \) (data augmentation)

(1) If \( y^{*} \) were observed, this would be a linear regression
\( (y_{i} \text{ would not be useful since it is just } \text{sgn}(y_{i}^{*}).) \)

We saw in the linear model before, \( p(\mathbf{\beta} \mid y_{i}^{*}, y_{i}) \)

(2) If (only) \( \mathbf{\beta} \) were observed, \( y_{i}^{*} \) would be a draw from
the normal distribution with mean \( \mathbf{x}_{i}^{\prime} \mathbf{\beta} \) and variance 1.
But, \( y_{i} \) gives the sign of \( y_{i}^{*} \). \( y_{i}^{*} \mid \mathbf{\beta}, y_{i} \) is a draw from
the truncated normal (above if \( y=0 \), below if \( y=1 \))
Gibbs Sampling for the Probit Model

(1) Choose an initial value for $\beta$ (maybe the MLE)
(2) Generate $y_i^*$ by sampling N observations from the truncated normal with mean $x_i'\beta$ and variance 1, truncated above 0 if $y_i = 0$, from below if $y_i = 1$.
(3) Generate $\beta$ by drawing a random normal vector with mean vector $(X'X)^{-1}X'y^*$ and variance matrix $(X'X)^{-1}$
(4) Return to 2 10,000 times, retaining the last 5,000 draws - first 5,000 are the 'burn in.'
(5) Estimate the posterior mean of $\beta$ by averaging the last 5,000 draws.
(This corresponds to a uniform prior over $\beta$.)

(This corresponds to a uniform prior over $\beta$.)
Generating Random Draws from f(X)

The inverse probability method of sampling random draws:
If F(x) is the CDF of random variable x, then a random draw on x may be obtained as F^{-1}(u) where u is a draw from the standard uniform (0,1).

Examples:
Exponential: f(x)=\theta exp(-\theta x); F(x)=1-exp(-\theta x)
\[ x = -(1/\theta)\log(1-u) \]
Normal: F(x) = \Phi(x); x = \Phi^{-1}(u)
Truncated Normal: x=\mu_i + \Phi^{-1}[1-(1-u)*\Phi(\mu_i)] for y=1;
\[ x= \mu_i + \Phi^{-1}[u\Phi(-\mu_i)] \text{ for } y=0. \]
Part 24: Bayesian Estimation

? Generate raw data
Calc ; Ran(13579) $
Sample ; 1 - 250 $
Create ; x1 = rnn(0,1) ; x2 = rnn(0,1) $
Create ; ys = .2 + .5*x1 - .5*x2 + rnn(0,1) ; y = ys > 0 $
Namelist; x = one,x1,x2$
Matrix ; xxx = <x'x> $
Calc ; Rep = 200 ; Ri = 1/(Rep-25)$
? Starting values and accumulate mean and variance matrices
Matrix ; beta=[0/0/0] ; bbar=init(3,1,0);bv=init(3,3,0)$$
Proc = gibbs $ Markov Chain – Monte Carlo iterations
Do for ; simulate ; r =1,Rep $
? ------- [ Sample y* | beta ] ------------------------------
Create ; mui = x'beta ; f = rnu(0,1)
  ; if(y=1) ysg = mui + inp(1-(1-f)*phi( mui));
     (else) ysg = mui + inp( f *phi(-mui)) $
? ------- [ Sample beta | y*] ------------------------------
Matrix ; mb = xxx*x'ysg ; beta = rndm(mb,xxi) $
? ------- [ Sum posterior mean and variance. Discard burn in. ]
Matrix ; if[r > 25] ; bbar=bbar+beta ; bv=bv+beta*beta'$
Enddo ; simulate $
Endproc $
Execute ; Proc = Gibbs $
Matrix ; bbar=ri*bbar ; bv=ri*bv-bbar*bbar' $
Probit ; lhs = y ; rhs = x $
Matrix ; Stat(bbar,bv,x) $
Example: Probit MLE vs. Gibbs

```r
--> Matrix ; Stat(bbar,bv); Stat(b,varb) $
+---------------------------------------------------+
|Number of observations in current sample = 1000 |
|Number of parameters computed here       = 3   |
|Number of degrees of freedom             = 997 |
+---------------------------------------------------+
+---------------------------------------------------+
|Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] |
|---------+-------------+----------------+----------+--------+
|BBAR_1   | .21483281    | .05076663       | 4.232    | .0000  |
|BBAR_2   | .40815611    | .04779292       | 8.540    | .0000  |
|BBAR_3   | -.49692480   | .04508507       | -11.022  | .0000  |
+---------------------------------------------------+
|Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] |
|---------+-------------+----------------+----------+--------+
|B_1      | .22696546    | .04276520       | 5.307    | .0000  |
|B_2      | .40038880    | .04671773       | 8.570    | .0000  |
|B_3      | -.50012787   | .04705345       | -10.629  | .0000  |
```
A Random Effects Approach

- Allenby and Rossi, “Marketing Models of Consumer Heterogeneity”
  - Discrete Choice Model – Brand Choice
  - “Hierarchical Bayes”
  - Multinomial Probit

- Panel Data: Purchases of 4 brands of Ketchup
Structure

Conditional data generation mechanism

\[ y_{it,j}^* = \beta_i x_{it,j} + \varepsilon_{it,j} , \] Utility for consumer \( i \), choice \( t \), brand \( j \).

\[ Y_{it,j} = 1[y_{it,j}^* = \text{maximum utility among the } J \text{ choices}] \]

\[ x_{it,j} = \text{(constant, log price, "availability," "featured")} \]

\[ \varepsilon_{it,j} \sim N[0, \lambda_j], \lambda_1 = 1 \]

Implies a \( J \) outcome multinomial probit model.
Bayesian Priors

Prior Densities

\[ \beta_i \sim N[\bar{\beta}, V_\beta], \]

Implies \( \beta_i = \bar{\beta} + w_i, w_i \sim N[0, V_\beta] \)

\( \lambda_j \sim Inverse \Gamma[v, s_j] \) (looks like chi-squared), \( v=3, s_i = 1 \)

Priors over model parameters

\[ \bar{\beta} \sim N[\bar{\beta}, aV_\beta], \bar{\beta} = 0 \]

\[ V_\beta^{-1} \sim Wishart[v_0, V_0], v_0 = 8, V_0 = 8I \]
Bayesian Estimator

- Joint Posterior \( E[\beta_1, \ldots, \beta_N, \bar{\beta}, V_\beta, \lambda_1, \ldots, \lambda_J \mid \text{data}] \)
- Integral does not exist in closed form.
- Estimate by random samples from the joint posterior.
- Full joint posterior is not known, so not possible to sample from the joint posterior.
Gibbs Cycles for the MNP Model

- Samples from the marginal posteriors
  - Marginal posterior for the individual parameters
    (Known and can be sampled)
    \[ \beta_i \mid \bar{\beta}, \mathbf{V}_\beta, \lambda, \text{data} \]
  - Marginal posterior for the common parameters
    (Each known and each can be sampled)
    \[ \bar{\beta} \mid \mathbf{V}_\beta, \lambda, \text{data} \]
    \[ \mathbf{V}_\beta \mid \bar{\beta}, \lambda, \text{data} \]
    \[ \lambda \mid \bar{\beta}, \mathbf{V}_\beta, \text{data} \]
Results

- Individual parameter vectors and disturbance variances
- Individual estimates of choice probabilities
- The same as the “random parameters model” with slightly different weights.
- Allenby and Rossi call the classical method an “approximate Bayesian” approach.
  - (Greene calls the Bayesian estimator an “approximate random parameters model”)
  - Who’s right?
    - Bayesian layers on implausible uninformative priors and calls the maximum likelihood results “exact” Bayesian estimators
    - Classical is strongly parametric and a slave to the distributional assumptions.
    - Bayesian is even more strongly parametric than classical.
    - Neither is right – Both are right.
Comparison of Maximum Simulated Likelihood and Hierarchical Bayes

- Ken Train: “A Comparison of Hierarchical Bayes and Maximum Simulated Likelihood for Mixed Logit”
- Mixed Logit

\[ U(i, t, j) = \beta'_i x(i, t, j) + \varepsilon(i, t, j), \]

\[ i = 1, \ldots, N \text{ individuals}, \]

\[ t = 1, \ldots, T_i \text{ choice situations} \]

\[ j = 1, \ldots, J \text{ alternatives (may also vary)} \]
Stochastic Structure – Conditional Likelihood

\[
\text{Prob}(i, j, t) = \frac{\exp(\beta_i'x_{i,j,t})}{\sum_{j=1}^{J} \exp(\beta_i'x_{i,j,t})}
\]

\[
\text{Likelihood} = \prod_{t=1}^{T} \frac{\exp(\beta_i'x_{i,j^*,t})}{\sum_{j=1}^{J} \exp(\beta_i'x_{i,j^*,t})}
\]

\(j^*\) = indicator for the specific choice made by \(i\) at time \(t\).

*Note individual specific parameter vector, \(\beta_i\)*
Classical Approach

\[ \beta_i \sim N[b, \Omega]; \text{ write } \Omega = \Gamma \Gamma' \]

\[ \beta_i = b + w_i \]

\[ = b + \Gamma v_i \text{ where } \Gamma = \text{diag}(\gamma_j^{1/2}) \text{ (uncorrelated)} \]

\[
\text{Log - likelihood} = \sum_{i=1}^{N} \log \int_w \prod_{t=1}^{T} \frac{\exp[(b + w_i)'x_{i,j*,t}]}{\sum_{j=1}^{J} \exp[(b + w_i)'x_{i,j,t}]} dw_i
\]

Maximize over \(b, \Gamma\) using maximum simulated likelihood
(random parameters model)
Bayesian Approach – Gibbs Sampling and Metropolis-Hastings

\[ \text{Posterior} = \prod_{i=1}^{N} L(\text{data} | \beta_i, \Omega) \times \text{priors} \]

\[ \text{Prior} = N(\beta_1, \ldots, \beta_N | b, \Omega) \text{ (normal)} \]
\[ \times IG(\gamma_1, \ldots, \gamma_N | \text{parameters}) \text{ (Inverse gamma)} \]
\[ \times g(b | \text{assumed parameters}) \text{ (Normal with large variance)} \]
Gibbs Sampling from Posteriors: \( b \)

\[
p(b | \beta_1, ..., \beta_N, \Omega) = \text{Normal}[\bar{\beta}, (1 / N)\Omega]
\]

\[
\bar{\beta} = (1 / N) \sum_{i=1}^{N} \beta_i
\]

Easy to sample from Normal with known mean and variance by transforming a set of draws from standard normal.
Gibbs Sampling from Posteriors: $\Omega$

\[
p(\gamma_k \mid b, \beta_1, ..., \beta_N) \sim \text{Inverse Gamma}[1 + N, 1 + N\bar{V}_k]
\]
\[
\bar{V}_k = \frac{1}{N} \sum_{i=1}^{N} (\beta_{k,i} - b_k)^2 \quad \text{for each } k=1, ..., K
\]

Draw from inverse gamma for each $k$:

Draw $1+N$ draws from $N[0,1] = h_{r,k}$,

then the draw is

\[
\frac{(1+N\bar{V}_k)}{\sum_{r=1}^{R} h_{r,k}^2}
\]
Gibbs Sampling from Posteriors: $\beta_i$

$$p(\beta_i \mid b, \Omega) = M \times L(data \mid \beta_i) \times g(\beta_i \mid b, \Omega)$$

$M=\text{a constant, } L=\text{likelihood, } g=\text{prior}$

(This is the definition of the posterior.)

Not clear how to sample.

Use Metropolis Hastings algorithm.
Metropolis – Hastings Method

**Define:**

\[ \beta_{i,0} = \text{an 'old' draw (vector)} \]

\[ \beta_{i,1} = \text{the 'new' draw (vector)} \]

\[ d_r = \sigma \Gamma v_r, \]

\[ \sigma = \text{a constant (see below)} \]

\[ \Gamma = \text{the diagonal matrix of standard deviations} \]

\[ v_r = \text{a vector of K draws from standard normal} \]
Metropolis Hastings: A Draw of $\beta_i$

\[ \text{Trial value: } \tilde{\beta}_{i,1} = \beta_{i,0} + d_r \]

\[ R = \frac{\text{Posterior}(\tilde{\beta}_{i,1})}{\text{Posterior}(\beta_{i,0})} \quad (Ms \ cancel) \]

\[ U = \text{a random draw from } U(0,1) \]

If $U < R$, use $\tilde{\beta}_{i,1}$, else keep $\beta_{i,0}$

During Gibbs iterations, draw $\beta_{i,1}$

$\sigma$ controls acceptance rate. Try for .4.
Application: Energy Suppliers

- $N=361$ individuals, 2 to 12 hypothetical suppliers
- $X =$ (1) fixed rates,  
  (2) contract length,  
  (3) local (0,1),  
  (4) well known company (0,1),  
  (5) offer TOD rates (0,1),  
  (6) offer seasonal rates (0,1).
# Estimates: Mean of Individual $\beta_i$

<table>
<thead>
<tr>
<th></th>
<th>MSL Estimate</th>
<th>Bayes Posterior Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>-1.04 (0.396)</td>
<td>-1.04 (0.0374)</td>
</tr>
<tr>
<td>Contract</td>
<td>-0.208 (0.0240)</td>
<td>-0.194 (0.0224)</td>
</tr>
<tr>
<td>Local</td>
<td>2.40 (0.127)</td>
<td>2.41 (0.140)</td>
</tr>
<tr>
<td>Well Known</td>
<td>1.74 (0.0927)</td>
<td>1.71 (0.100)</td>
</tr>
<tr>
<td>TOD</td>
<td>-9.94 (0.337)</td>
<td>-10.0 (0.315)</td>
</tr>
<tr>
<td>Seasonal</td>
<td>-10.2 (0.333)</td>
<td>-10.2 (0.310)</td>
</tr>
</tbody>
</table>
Reconciliation: A Theorem (Bernstein-Von Mises)

- The posterior distribution converges to normal with covariance matrix equal to $1/n$ times the information matrix (same as classical MLE). (The distribution that is converging is the posterior, not the sampling distribution of the estimator of the posterior mean.)
- The posterior mean (empirical) converges to the mode of the likelihood function. Same as the MLE. A proper prior disappears asymptotically.
- Asymptotic sampling distribution of the posterior mean is the same as that of the MLE.