MODERN BAYESIAN ECONOMETRICS

LECTURES

BY TONY LANCASTER

January 2006

AN OVERVIEW

These lectures are based on my book

An Introduction to Modern Bayesian Econometrics,

Blackwells, May 2004 and some more recent material.

The main software used is WinBUGS http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

This is shareware.

Practical classes using WinBUGS accompany these lectures.

The main programming and statistical software is R.

http://www.r-project.org/

This is also shareware.

There is also R to Matlab connectivity – see the r-project home page.

Also see BACC Bayesian econometric software – link on the course web page.

These introductory lectures are intended for both econometricians and applied economists in general.

Rev. T. Bayes

Figure 1:

AIM

The aim of the course is to explain how to do econometrics the Bayesian way.

Rev. Thomas Bayes (1702-1761)

METHOD

By computation.

Dominant approach since 1990.

Superceding earlier heavy algebra.

OUTLINE

Principles of Bayesian Inference

Examples

Bayesian Computation and MCMC

\

PRINCIPLES (Chapter 1)

Bayes theorem for events:

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

Bayes' theorem for densities:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Bayes theorem for parameters and data:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{2}$$

Notation for data — y or y^{obs}.

So Bayes theorem transforms prior or initial probabilities, $\Pr(A)$, into posterior or subsequent probabilities, $\Pr(A|B)$.

B represents some new evidence or data and the theorem shows how such evidence should change your mind.

EXAMPLES OF BAYES THEOREM

(with possible, and debatable, likelihoods and priors)

1. Jeffreys' Tramcar Problem

Trams are numbered $1, 2, 3, ...n$. A stranger (Thomas Bayes?) arrives at the railway station and notices tram number m. He wonders how many trams the city has.

$$p(n|m) = \frac{p(m|n)p(n)}{p(m)} \propto p(m|n)p(n)$$

Jeffreys' solution: Take $p(n) \propto 1/n$ and $p(m|n) = 1/n$ – i.e. uniform. Then

$$p(n|m) \propto \frac{1}{n^2} \quad n \geq m$$

strictly decreasing with median (about) $2m$. A reasonable guess if he sees tram 21 might therefore be 42.

2 A Medical Shock

A rare but horrible disease D or its absence $\overline{D}$.

A powerful diagnostic test with results $+$ (!) or $-$.

$$
\begin{aligned}
\Pr(D) &= 1/10000 \quad \text{(rare)} \\
\Pr(+|D) &= 0.9 \quad \text{(powerful test)} \\
\Pr(+|\overline{D}) &= 0.1\text{....(false positive)} \\
\Pr(D|+) &= \frac{\Pr(+|D)\Pr(D)}{\Pr(+)} \\
&= \frac{\Pr(+|D)\Pr(D)}{\Pr(+|D)\Pr(D) + \Pr(+|\overline{D})\Pr(\overline{D})} \\
&= \frac{0.90}{0.90 + 0.10(10,000-1)} \sim \frac{0.9}{1000} \\
&= 0.0009 \quad \text{(relief)}
\end{aligned}
$$

3. Paradise Lost?[1]

If your friend read you her favourite line of poetry and told you it was line $(2, 5, 12, 32, 67)$ of the poem, what would you predict for the total length of the poem?

Let l be total length and y the length observed. Then by Bayes theorem

$$p(l|y) \propto p(y|l)p(l)$$

Take $p(y|l) \propto 1/l$ (uniform) and $p(l) \propto l^{-\gamma}$. Then

$$p(l|y) \propto l^{-(1+\gamma)}, \quad l \geq t \qquad (*)$$

The density $p(y|l)$ captures the idea that the favourite line is equally likely to be anywhere in the poem; the density $p(l)$ is empirically roughly accurate for some γ.

Experimental subjects asked these (and many similar) questions reply with predictions consistent with the median of (*)

[1] *Optimal predictions in everyday cognition*,Griffiths and Tenenbaum, forthcoming in Psychological Science.

INTERPRETATION OF Pr(.)

Probability as rational degree of belief in a proposition.

Not "limiting relative frequency". Not "equally likely cases".

Ramsey "Truth and Probability" (1926)

See the web page for links to Ramsey's essay

Persi Diaconis. "Coins don't have probabilities, people do". "Coins don't have little numbers P hidden inside them."

Later, deFinetti. "Probability does not exist".

Let θ be the parameter of some economic model and let y be some data.

Prior is

$$p(\theta)$$

Likelihood is

$$p(y|\theta)$$

Marginal Likelihood or Predictive Distribution of the (potential) data is

$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

.

Posterior Distribution is

$$p(\theta|y)$$

.

The Bayesian Algorithm (page 9)

1. Formulate your economic model as a collection of probability distrbutions conditional on different values for a parameter θ, about which you wish to learn.

2. Organize your beliefs about θ into a (prior) probability distribution.

3. Collect the data and insert them into the family of distributions given in step 1.

4. Use Bayes' theorem to calculate your new beliefs about θ.

5. Criticise your model.

The Evolution of Beliefs

Consider the following data from 50 Bernoulli trials

0 0 1 0 0 1 0 0 0 0 0 1 0 1 1 1 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0

If θ is the probability of a one at any one trial then the likelihood of any sequence of s trials containing y ones is

$$p(y|\theta) = \theta^y(1-\theta)^{s-y}$$

So if the prior is uniform – $p(\theta) = 1$ – then after 5 trials the posterior is

$$p(\theta|y) \propto \theta(1-\theta)^4$$

and after 10 trials the posterior is

$$p(\theta|y) \propto \theta(1-\theta)^4 \times \theta(1-\theta)^4 = \theta^2(1-\theta)^8$$

and after 40 trials the posterior is

$$p(\theta|y) \propto \theta^2(1-\theta)^8\theta^8(1-\theta)^{22} = \theta^{10}(1-\theta)^{30}$$

and after all 50 trials the posterior is

$$p(\theta|y) = \theta^{10}(1-\theta)^{30}\theta^4(1-\theta)^6 = \theta^{14}(1-\theta)^{36}$$

These successive posteriors are plotted below

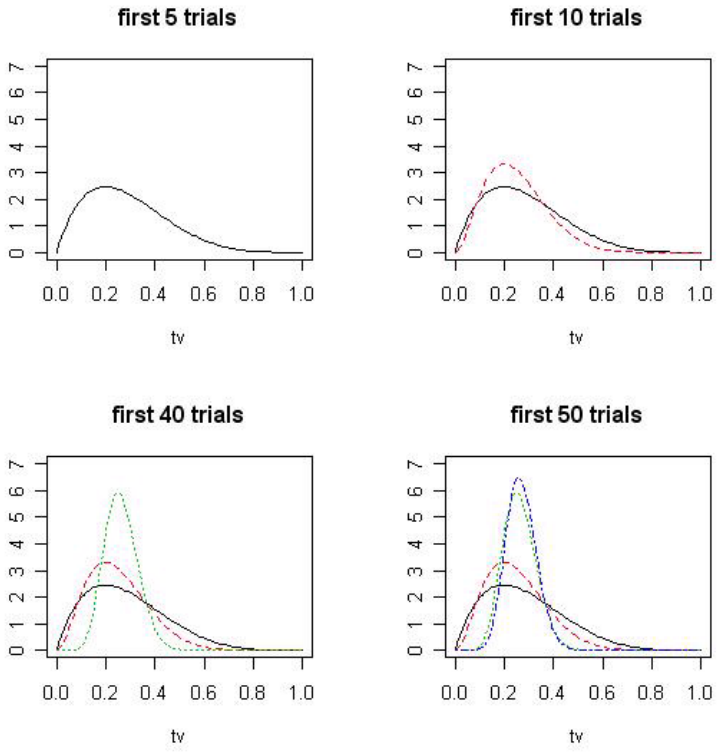


Figure 2:

Note:

1. The previous posterior becomes the new prior

2. Beliefs seem to become more concentrated as the number of observations increases.

3. Posteriors seem to look more normal as the number of observations increases.

4. (Not shown here) The prior has less and less influence on the posterior as $n \to \infty$.

These are quite general properties of posterior inference.

Proper and Improper Priors (section 1.4.2)

Any proper probability distribution over $\theta \in \Theta$ will do.

Bayes' theorem shows only how beliefs change. It does not dictate what beliefs should be.

The theorem shows how beliefs are changed by evidence.

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

A model for the evolution of scientific knowledge?

Improper priors are sometimes used. These do not integrate to one and are not probability distributions. Simple examples are

$$p(\beta) \propto 1, \qquad -\infty < \beta < \infty \tag{3}$$

$$p(\sigma) \propto 1/\sigma \qquad \sigma > 0 \tag{4}$$

Can be thought of as approximations to diffuse but proper priors.

What matters is that the posterior is proper. e.g. in the tramcar problem the prior $1/n$ was improper but the posterior $1/n^2$, $n \geq m$ was proper.

Improper priors sometimes mathematically convenient. But software e.g. WinBUGS requires proper priors.

Notice use of $\propto$ which means "is proportional to". Scale factors irrelevant in most Bayesian calculation.

Some people want "objective" priors that can be generated by applying a rule.

In particular there is a desire for a rule that can generate "non-informative" priors.

Others are content to form priors subjectively and then to study the effect, if any, of changing them.

There are several general rules which I'll mention fairly briefly. They all have drawbacks.

One rather important rule (which doesn't have a drawback) is:

Don't assign probability zero to parts of the parameter space.

This is because the posterior is the product of likelihood and prior so prior probability zero $\Rightarrow$ posterior probability zero. So you can't learn that you were wrong.

Natural Conjugate Priors(pps 30-33)

Class of priors is conjugate for a family of likelihoods if both prior and posterior are in the same class for all data y.

Natural Conjugate prior has the same functional form as the likelihood e.g.Bernoulli likelihood

$$\ell(\theta; y) \propto \theta^y (1-\theta)^{1-y} \tag{5}$$

and Beta prior

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1} \tag{6}$$

giving (Beta) posterior

$$p(\theta|y) \propto \theta^{y+a-1}(1-\theta)^{b-y}. \tag{7}$$

Note symbol ℓ to represent likelihood previously described by $p(y|\theta)$. Terminology of R. A. Fisher.

Natural conjugate can be thought of as a posterior from (hypothetical) previous sample. So posterior stays in the same family but its parameters change. Quite convenient if you're doing algebra but of little relevance if you're computing. Conjugate form sometimes conflicts with reasonable prior beliefs.

Jeffreys' Priors (pps 34-37)

Fisher information about θ is

$$\mathcal{I}(\theta) = -E\frac{\partial^2 \log \ell(\theta; y)}{\partial \theta^2}|\theta \tag{8}$$

Jeffreys' prior is $\propto |\mathcal{I}(\theta)|^{1/2}$.

This prior is *invariant* to reparametrization. Posterior beliefs about θ are the same whether prior expressed on θ or on $\phi(\theta)$.

Jeffreys' priors can also be shown to in a certain sense minimally informative relative to the likelihood.

Example:: iid normal data mean zero precision τ

$$\begin{aligned} \ell(\tau; y) &\propto \tau^{n/2} \exp\{-\tau \Sigma_{i=1}^{n} y_i^2/2\} \\ \mathcal{I}(\theta) &= \frac{n}{2\tau^2} \\ \text{so } p(\theta) &\propto \frac{1}{\tau}. \end{aligned}$$

Subjective Priors

Economic agents have subjective priors but for econometricians ?

Econometric modelling arguably subjective.

Arguments that an instrumental variable is "valid" typically subjective.

Why is randomized allocation of treatments convincing?

Can always study sensitivity of inferences to changes in the prior.

Hierarchical Priors (pages 37-39)

Often useful to think about the prior for a vector parameters in stages.

Suppose that $\theta = (\theta_1, \theta_2, ...\theta_n)$ and λ is a parameter of lower dimension than θ. Then to get $p(\theta)$ consider $p(\theta|\lambda)p(\lambda) = p(\theta, \lambda)$ so that $p(\theta) = \int p(\theta|\lambda)p(\lambda)d\lambda$. λ is a hyperparameter. And

$$p(\theta, \lambda, y) = p(y|\theta)p(\theta|\lambda)p(\lambda).$$

Example: $\{y_i\} \sim n(\theta_i, \tau)$; $\{\theta_i\} \sim n(\mu, \phi)$; $p(\phi) \propto 1$ then the posterior means of θ_i take the form

$$E(\theta_i|y_1, ...y_n) = \frac{\tau y_i + \phi \overline{y}}{\tau + \phi}$$

Example of shrinkage to the general mean. c.f. estimation of permanent incomes.

Any proper probability distribution for Y will do.

Can always study sensitivity of inferences to changes in the likelihood.

Posterior Distributions(pps 41 - 55)

Express what YOU believe given model and data.

Parameter θ and data y are usually vector valued.

Interest often centres on individual elements, e.g. θ_i. The posterior distribution of θ_i

$$p(\theta_i|y) = \int p(\theta|y)d\theta_{(-i)} \tag{9}$$

Bayesian methods involve integration.

This was a barrier until recently. No longer

For example, WinBUGS does high dimensional numerical integration.

NO reliance on asymptotic distribution theory – Bayesian results are "exact".

Frequentist (Classical) Econometrics (Appendix 1)

Relies (mostly) on distributions of estimators and test statistics over hypothetical repeated samples. Does NOT condition on the data. Inferences are based on data not observed!

Such sampling distributions are strictly irrelevant to Bayesian inference.

Sampling distributions arguably arbitrary e.g. fixed versus random regressors. Conditioning an ancillaries.

In Bayesian work there are no estimators and no test statistics.

So there is no role for unbiasedness, minimum variance, efficiency etc.

"p values" give probability of the data given the hypothesis. Reader wants probability of the hypothesis given the data.

Probability of bloodstain given guilty .V. Probability guilty given bloodstain! Prosecutor's fallacy.

One Parameter Likelihood Examples

1. Normal regression of consumption, c, on income, y. (pps 12-13)

$$\ell(\beta; c, y) \;\propto\; \exp\{-(\tau\Sigma y_i^2/2)(\beta - b)^2\}$$

$$\text{for} \quad b \;=\; \Sigma_{i=1}^n c_i y_i / \Sigma_{i=1}^n y_i^2.$$

(The manipulation here was

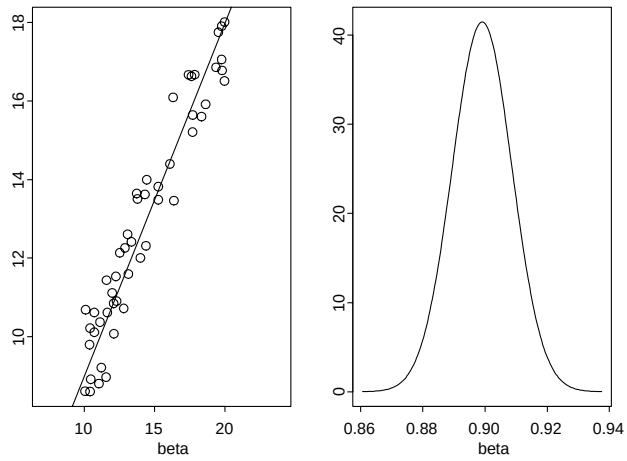
$$\Sigma(c - \beta y)^2 = \Sigma(c - by + (b - \beta)y)^2 = \Sigma e^2 + (\beta - b)^2 \Sigma y^2)$$

Note notation: ℓ for likelihood; $\propto$ for "is proportional to"; τ for precision – $1/\sigma^2$.

So β is normally distributed.

Likelihood has the shape of a normal density with mean b and precision $\tau\Sigma_i y_i^2$.

Figure 3: Plot of the Data and the Likelihood for Example 1



2. Autoregression (pps 14-16)

$$p(y|y_1, \rho) \propto \exp\{-(\tau/2)\Sigma_{t=2}^{T}(y_t - \rho y_{t-1})^2\}.$$

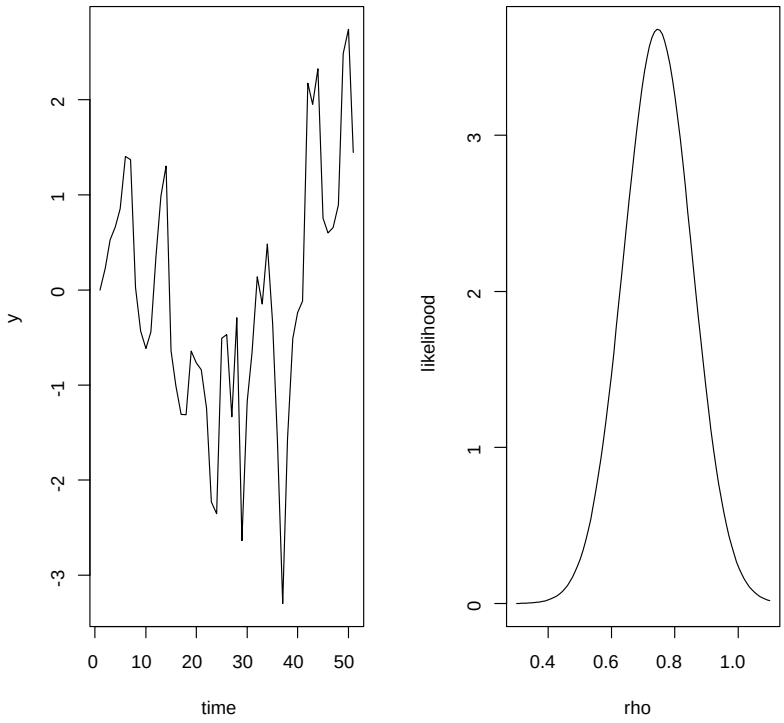
Rearranging the sum of squares in exactly the same way as in example 1 and then regarding the whole expression as a function of ρ gives the likelihood kernel as

$$\ell(\rho; y, y_1, \tau) \;\propto\; \exp\{-(\tau\Sigma_{t=2}^{T} y_{t-1}^2/2)(\rho - r)^2\}$$

$$\text{for} \quad r \;=\; \Sigma_{t=2}^{T} y_t y_{t-1} / \Sigma_{t=2}^{T} y_{t-1}^2.$$

Note terminology: "kernel" of a density neglects multiplicative terms not involving the quantity of interest.

Figure 4: Time Series Data and its Likelihood



So ρ is normally distributed (under a uniform prior).

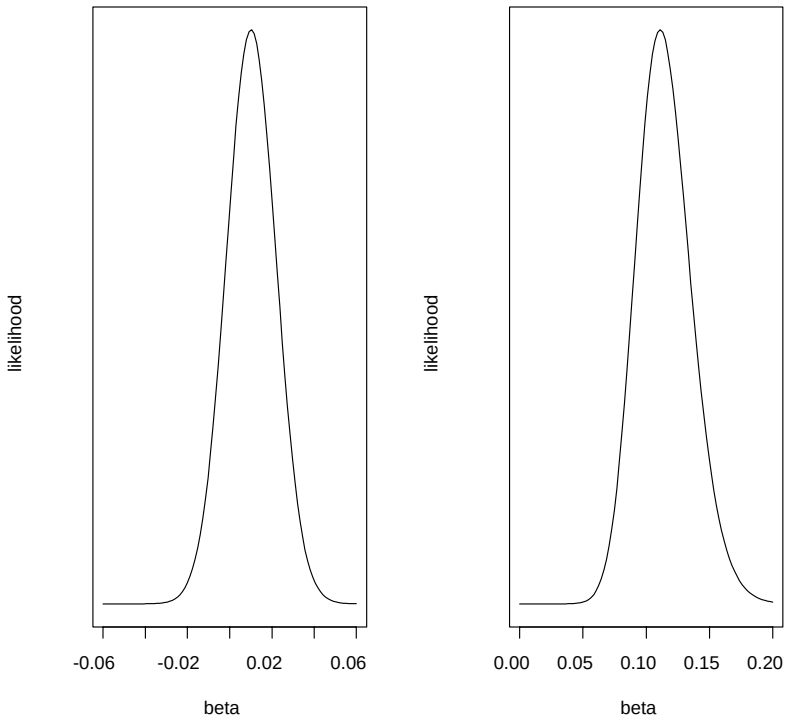
3. Probit model (pps 17-18)

$$\ell(\beta; y, x) = \Pi_{i=1}^{n} \Phi(\beta x_i)^{y_i} (1 - \Phi(\beta x_i))^{1-y_i}.$$

Figures for $n = 50$. $\beta = 0$. Simulated data with $\beta = 0$ for fig 1 and $\beta = 0.1$ for fig 2.a

For both likelihoods the function is essentially zero everywhere else on the real line!

Figure 5: Two Probit Likelihoods



4. Example Laplace data: (pps 61-63)

$$p(y|\theta) = \exp\{-|y-\theta|\}, \qquad -\infty < y, \theta < \infty.$$

Thick tailed compared to normal. Figure plots the Laplace density function for the case $\theta = 1$.

Figure 6: A Double Exponential Density

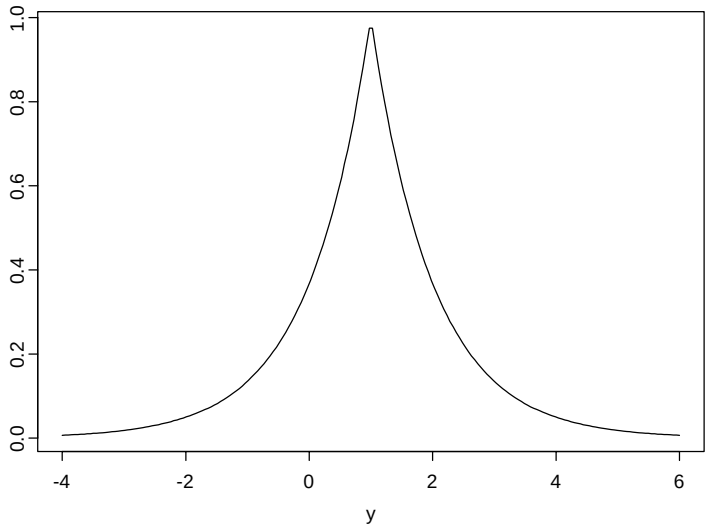
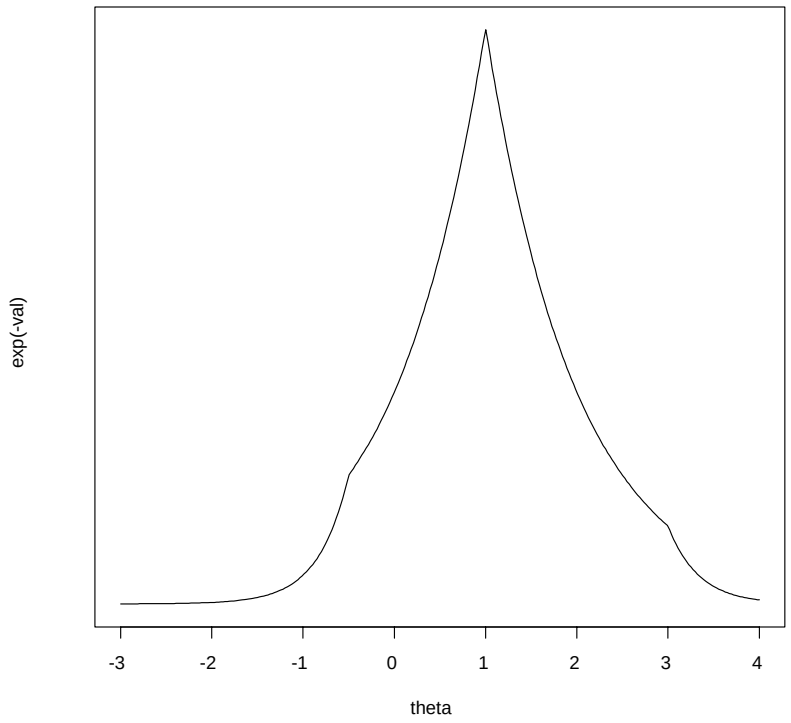


Figure 7: The Likelihood for 3 Observations of a Laplace Variate



A Nonparametric (Multinomial) Likelihood (pps 141-147)

$$\Pr(Y = y_l) = p_l \tag{10}$$

$$\ell(p; y) \propto \Pi_{l=1}^{L} p_l^{n_l}. \tag{11}$$

Natural conjugate prior for $\{p_i\}$ is the Dirichlet (multivariate Beta).

$$p(p) \propto \Pi_{l=0}^{L} p_l^{\nu_l - 1}$$

Posterior can be simulated as $p_l = g_l / \Sigma_{i=1}^{L} g_i$ where $\{g_i\} \sim$ iid unit Exponential as $\{\nu_l\} \to 0$.

L may be arbitrarily large.

Since as $\{\nu_l\} \to 0$ the posterior density of the $\{p_l\}$ concentrate on the observed data points the posterior density of, say,

$$\mu = \Sigma_{l=1}^{l} p_l y_l \tag{12}$$

– difficult to find analytically – may be easily found by simulation as

$$\mu = \frac{\Sigma_{i=1}^{n} y_i g_i}{\Sigma_{i=1}^{n} g_i}, \qquad \{g_i\} \sim iid \text{ E}(1). \tag{13}$$

For example

```
g <- rexp(n);
```

```
mu <- sum(g*y)/sum(g).
```

Equation (12) is a moment condition. This is a Bayesian version of method of moments. (We'll give another later.) Also called Bayesian Bootstrap.

To see why this called a bootstrap and the precise connection with the frequentist bootstrap see my paper *A Note on Bootstraps and Robustness* on the web site.

What is a parameter? (pps 21-22)

Anything that isn't data.

Example: Number of tramcars.

Example: How many trials did he do?

n Bernoulli trials with a parameter θ agreed to be 0.5. $s = 7$, successes recorded. What was n? The probability of s successes in n Bernoulli trials is the binomial expression

$$P(S = s|n, \theta) = \binom{n}{s}\theta^s(1-\theta)^{n-s}, \qquad s = 0, 1, 2, ...n, \ 0 \leq \theta \leq 1, \tag{14}$$

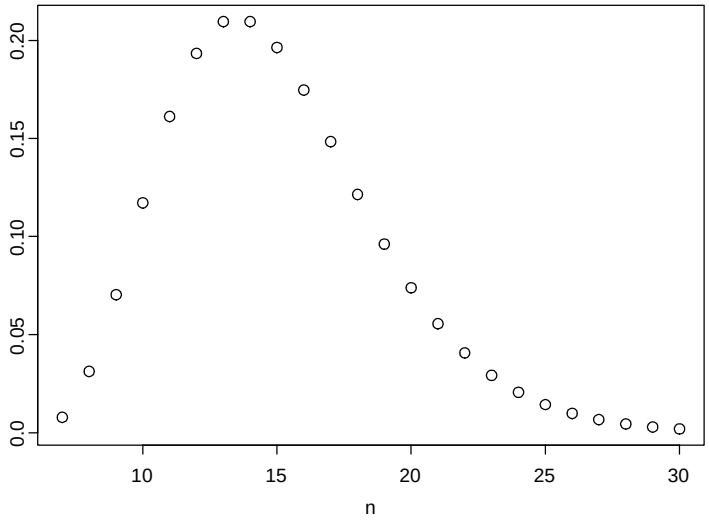
and on inserting the known data $s = 7, \theta = 1/2$ we get the likelihood for the parameter n

$$\ell(n; s, \theta) \propto \frac{n!}{(n-7)!}\left(\frac{1}{2}\right)^n, \qquad n \geq 7.$$

This is drawn in the next figure for $n = 7, 8, ...30$.

Mode at $2n$ of course.

Figure 8: Likelihood for n



Another example: Which model is true? The label of the true! model is a parameter. It will have a prior distribution and, if data are available, it will have a posterior distribution.

Inferential Uses of Bayes' Theorem

Bayesian inference is based entirely upon the (marginal) posterior distribution of the quantity of interest.

"Point Estimation"

Posterior mode(s), mean etc.

Or *decision theory* perspective. (pps 56-57) Minimize $\int \text{loss}(\widehat{\theta}, \theta) p(\theta|y) d\theta$ – expected posterior loss w.r.t $\widehat{\theta}$. Quadratic loss

$$\text{loss}(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$$

leads to the posterior mean.

Absolute error loss

$$\text{loss}(\widehat{\theta}, \theta) = |\widehat{\theta} - \theta|$$

leads to the posterior median.

Example: Probit model. Suppose the parameter of interest is

$$\partial P(y = 1|x, \beta)/\partial x_j$$

at $x = \overline{x}$. This is a function of β. So compute its marginal posterior distribution and report the mean etc.

Example: Bernoulli trials: Assume (natural conjugate) beta family $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1},\ 0 \leq \theta \leq 1$. With data from n Bernoulli trials posterior is

$$p(\theta|y) \propto \theta^{s+a-1}(1-\theta)^{n-s+b-1}$$

with mean and variance

$$\begin{aligned} E(\theta|y) &= \frac{s+a}{n+a+b}, \\ V(\theta|y) &= \frac{(s+a)(n-s+b)}{(n+a+b)^2(n+a+b+1)}. \end{aligned}$$

For large n and s, n in the ratio r then approximately

$$E(\theta|y) = r, \qquad V(\theta|y) = \frac{r(1-r)}{n}.$$

Notice asymptotic irrelevance of the prior (if it's NOT dogmatic). This is a general feature of Bayesian inference. Log likelihood $O(n)$ but prior of $O(1)$.

Example: Maximum likelihood. Since $p(\theta|y) \propto \ell(\theta; y)p(\theta)$ ML gives the vector of (joint) posterior modes under a uniform prior. This differs, in general, from the vector of marginal modes or means.

Uniform Distribution (p 57)

Let Y be uniformly distributed on 0 to θ so

$$p(y|\theta) = \begin{cases} 1/\theta & \text{for } 0 \leq y \leq \theta \\ 0 & \text{elsewhere} \end{cases} \tag{15}$$

with likelihood for a random sample of size n

$$\ell(\theta; y) \propto \begin{cases} 1/\theta^n & \text{for } y_{\max} \leq \theta \\ 0 & \text{elsewhere} \end{cases} \tag{16}$$

Maximum likelihood estimator of θ is $y_{\max}$ which is always too small!

Bayes posterior expectation under prior $p(\theta) \propto 1/\theta$ is

$$E(\theta|y) = \frac{n}{n-1} y_{\max}. \tag{17}$$

"Interval Estimation" (p 43)

Construct a 95% highest posterior density interval (region). This is a set whose probability content is 0.95 and such that no point outside it has higher posterior density than any point inside it.

Example: $\Pr(\overline{x} - 1.96\sigma/\sqrt{n} < \mu < \overline{x} + 1.96\sigma/\sqrt{n}) = 0.95$ when data are iid $n(\mu, \sigma^2)$ with σ^2 known. This statement means what it says! It does not refer to hypothetical repeated samples.

For vector parameters construct highest posterior density regions.

Prediction (pps 79-97)

(i) of data to be observed.

Use $p(y) = \int p(y|\theta)p(\theta)d\theta$

(ii) of new data $\widetilde{y}$ given old data.

Use $p(\widetilde{y}|y) = \int p(\widetilde{y}|y,\theta)p(\theta|y)d\theta$

Example: Prediction from an autoregression with τ known and equal to one.

$$p(\widetilde{y}|y^{obs},\rho) \propto \exp\{-(1/2)(y_{n+1}-\rho y_n)^2\}$$

Thus, putting $s^2 = \Sigma_{t=2}^{n} y_{t-1}^2$, and using the fact established earlier that the posterior density of ρ is normal with mean r and precision s^2,

$$\begin{aligned} p(y_{n+1}|y) &\propto \int \exp\{-(1/2)(y_{n+1}-\rho y_n)^2 - (s^2/2)(\rho - r)^2\} d\rho \\ &\propto \exp\{-\frac{1}{2}\left(\frac{s^2}{s^2+y_n^2}\right)(y_{n+1}-ry_n)^2\} \end{aligned}$$

which is normal with mean equal to ry_n and precision $s^2/(s^2+y_n^2) < 1$.

$p(y_{n+1}|y)$ is the *predictive density* of y_{n+1}.

Prediction and Model Criticism (chapter 2)

$p(y)$ says what you think the data should look like.

You can use it to check a model by

1. Choose a "test statistic", $T(y)$
2. Calculate its predictive distribution from that of y
3. Find $T(y^{\text{obs}})$ and see if it is probable or not.

Step 2 can be done by sampling:

1. Sample θ from $p(\theta)$
2. Sample y from $p(y|\theta)$ and form $T(y)$
3. Repeat many times.

Model Choice(pps 97-102)

Let M_j denote the j'th of J models and let the data be y. Then by Bayes' theorem the posterior probability of this model is

$$\begin{aligned} P(M_j|y) &= \frac{p(y|M_j)P_j}{p(y)}, \\ \text{where } p(y) &= \Sigma_{j=1}^{J} p(y|M_j)P_j. \end{aligned}$$

and, with $J = 2$, the posterior odds on model 1 are

$$\frac{P(M_1|y)}{P(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)}\frac{P(M_1)}{P(M_2)}.$$

$p(y|M_j)$ are the predictive distributions of the data on the two hypotheses and their ratio is the *Bayes factor*.

For two simple hypotheses

$$\frac{P(\theta = \theta_1|y^{obs})}{P\left(\theta = \theta_2|y^{obs}\right)} = \frac{\ell(\theta_1; y^{obs})}{\ell(\theta_2; y^{obs})} \frac{P(\theta = \theta_1)}{P(\theta = \theta_2)}$$

In general the probability of the data given model j is

$$P(y|M_j) = \int \ell(y|\theta_j) p(\theta_j) d\theta_j \tag{18}$$

where $\ell(y|\theta_j)$ is the likelihood of the data under model j.

.

Example with Two Simple Hypotheses

$\ell(y;\theta)$ is the density of a conditionally normal $(\theta, 1)$ variate.

Two hypotheses are that $\theta = -1$ and $\theta = 1$ and sample size is $n = 1$. The likelihood ratio is

$$\frac{P(y^{obs}|\theta = -1)}{P\left(y^{obs}|\theta = 1\right)} = \frac{e^{-(1/2)(y+1)^2}}{e^{-(1/2)(y-1)^2}}$$

and so, if the hypotheses are equally probable a priori, the posterior odds are

$$\frac{P(\theta = -1|y^{obs})}{P\left(\theta = 1|y^{obs}\right)} = e^{-2y}.$$

If $y > 0$ then $\theta = 1$ more probable than $\theta = -1$; $y < 0$ makes $\theta = -1$ more probable than $\theta = 1$; $y = 0$ equal to zero leaves the two hypotheses equally probable

If you observe $y = 0.5$ then posterior odds on $\theta = 1$ are $e = 2.718$ corresponding to a probability of this hypothesis of $P(\theta = 1|y = 0.5) = e/(1 + e) = 0.73$. When $y = 1$ the probability moves to 0.88.

In the linear model an approximate Bayes factor is the BIC – Bayesian Information Criterion. The approximate Bayes factor in favour of model 2 compared to model 1 takes the form

$$BIC = \left(\frac{R_1}{R_2}\right)^{n/2} n^{(k_1-k_2)/2} \tag{19}$$

where the R_j are the residual sums of squares in the two models and the k_j are the numbers of coefficients.

For example

$$\text{Model 1} \qquad y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon_1 \tag{20}$$

$$\text{Model 2} \qquad y = \gamma_1 x_1 + \gamma_2 x_3 + \varepsilon_2 \tag{21}$$

Model Averaging

For prediction purposes one might not want to use the most probable model. Instead it is optimal, for certain loss functions, to predict from an average model using

$$\begin{aligned} p(\widetilde{y}|y) &= \Sigma_j p(\widetilde{y}, M_j|y) \\ &= \Sigma_j P(M_j|y) p(\widetilde{y}|M_j, y). \end{aligned}$$

So predictions are made from a weighted average of the models under consideration with weights provided by the posterior model probabilities

Linear Models (Chapter 3)

Normal linear model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim n(0, \tau I_n) \tag{22}$$

and conventional prior

$$p(\beta, \tau) \propto 1/\tau \tag{23}$$

yields

$$p(\beta|\tau, y, X) = n(b, \tau X'X) \tag{24}$$

$$p(\tau|y, X) = \text{gamma}(\frac{n-k}{2}, \frac{e'e}{2}) \tag{25}$$

where

$$b = (X'X)^{-1}X'y \quad \text{and} \quad e = y - Xb. \tag{26}$$

Marginal posterior density of β is multivariate t.

BUT the simplest way is to sample β, τ.

Algorithm:

1. Sample τ using rgamma

2. Put τ in to (20) and sample using mvrnorm.

3. Repeat 10,000 times.

This makes it easy to study the marginal posterior distribution of ANY function of β, τ.

A Non-Parametric Version of the Linear Model (pps 141-147)

(Bayesian Bootstrap Again)

Consider the linear model again but without assuming normality or homoscedasticity. Define β by

$$EX'(y - X\beta) = 0.$$

So,

$$\beta = [E(X'X)]^{-1}E(X'y)$$

Assume the rows of $(y : X)$ are multinomial with probabilities $p = (p_1, p_2, ...p_L)$. So a typical element of $E(X'X)$ is $\Sigma_{i=1}^{n} x_{il}x_{im}p_i$ and a typical element of $E(X'y)$ is $\Sigma_{i=1}^{n} x_{il}y_i p_i$. Thus we can write β as

$$\beta = (X'PX)^{-1}X'Py.$$

where $P = diag\{p_i\}$. If the prior for $\{p_i\}$ is Dirichlet (multivariate beta) then so is the posterior (natural conjugate) and, as before, the $\{p_i\}$ can be simulated by

$$p_i = \frac{g_i}{\Sigma_{j=1}^{n} g_j} \quad \text{for } i = 1, 2, ...n. \tag{27}$$

where the $\{g_i\}$ are independent unit exponential variates. So we can write

$$\beta \cong (X'GX)^{-1}X'Gy \tag{28}$$

where G is an $n \times n$ diagonal matrix with elements that are independent gamma(1), or unit exponential, variates. The symbol $\widetilde{=}$ means "is distributed as".

β has (approximate) posterior mean equal to the least squares estimate $b = (X'X)^{-1}X'y$ and its approximate covariance matrix is

$$V = (X'X)^{-1}X'DX(X'X)^{-1}; \quad D = diag\{e_i^2\},$$
$$\text{where } e = y - Xb.$$

This posterior distribution for β is the Bayesian bootstrap distribution. It is robust against heteroscedasticity and non-normality.

Can do (bb) using weighted regression with weights equal to `rexp(n)` – see exercises.

Example: Heteroscedastic errors and two real covariates: $n = 50$.

coefficient	ols	se	BB mean	White se	BB se
b_0	.064	.132	.069	.128	.124
b_1	.933	.152	.932	.091	.096
b_2	-.979	.131	-.974	.134	.134

Bayesian Method of Moments (Again) (not in book)

Entropy

Entropy measures the amount of uncertainy in a probability distribution. The larger the entropy the more the uncertainy. For a discrete distribution with probabilities $p_1, p_2, ...p_n$ entropy is

$$-\Sigma_{i=1}^{n} p_i \log p_i.$$

This is maximized subject to $\Sigma_{i=1}^{n} p_i = 1$ by $p_1 = p_2 = ... = p_n = 1/n$ which is the most uncertain or least informative distribution.

Suppose that all you have are moment restrictions of the form $Eg(y,\theta) = 0$. But Bayesian inference needs a likelihood. One way to proceed – Schennach, Biometrika 92(1), 2005 – is to construct a maximum entropy distribution supported on the observed data. This gives probability p_i to observation y_i. As we have seen the unrestricted maxent distribution assigns probability $1/n$ to each data point which is the solution to

$$\max_p \Sigma_{i=1}^n - p_i \log p_i \text{ subject to } \Sigma_{i=1}^n p_i = 1$$

The general procedure solves the problem

$$\max_p \Sigma_{i=1}^n - p_i \log p_i \text{ subject to } \Sigma_{i=1}^n p_i = 1 \text{ and } \Sigma_{i=1}^n p_i g(y_i,\theta) = 0 \tag{29}$$

The solution has the form

$$p_i^*(\theta) = \frac{\exp\{\lambda(\theta)' g(y_i,\theta)\}}{\Sigma_{j=1}^n \exp\{\lambda(\theta)' g(y_i,\theta)\}}$$

where the $\{\lambda(\theta)\}$ are the Lagrange multipliers associated with the moment constraints. The resulting posterior density takes the form

$$p(\theta|Y) = p(\theta)\Pi_{i=1}^n p_i^*(\theta)$$

where $p(\theta)$ is an arbitrary prior.

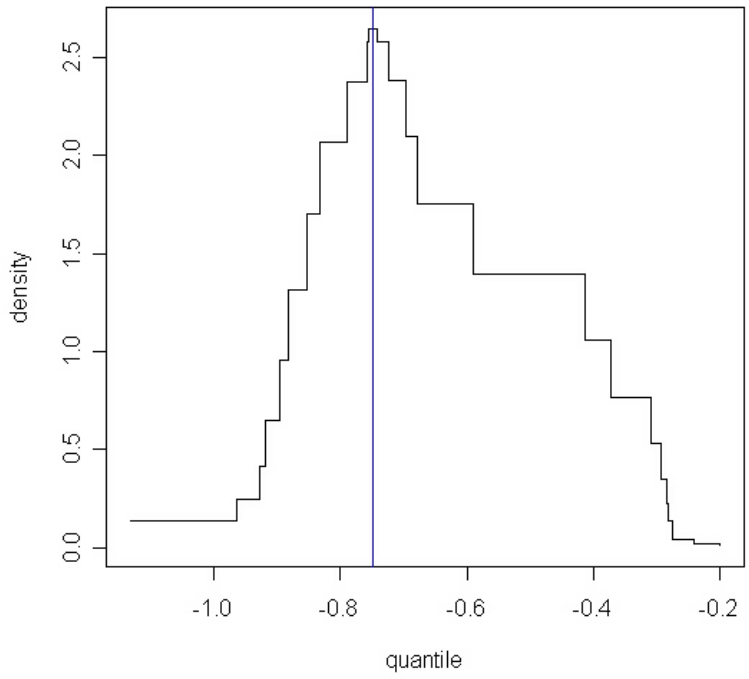


Figure 9:

Here is an example:

Estimation of the 25% quantile

Use the single moment

$$1(y \leq \theta) - 0.25$$

which has expectation zero when θ is the 25% quantile. Figure 7 shows the posterior density of the 25% quantile based on a sample of size 100 under a uniform prior. The vertical line is the sample 25% quantile.This method extends to **any** GMM setting including linear and non-linear models, discrete choice, instrumental variables etc.

BAYESIAN COMPUTATION AND MCMC(pps 183-192)

When the object of interest is, say $h(\theta)$ a scalar or vector function of θ Bayesian inferences are based on the marginal distribution of h. How do we obtain this?

The answer is the sampling principle, (just illustrated in the normal linear model and on several other occasions) that underlies all modern Bayesian work.

Sampling Principle: To study $h(\theta)$ sample from $p(\theta|y)$ and for each realization θ^i form $h(\theta^i)$. Many replications will provide, exactly, the marginal posterior density of h.

Example using R Suppose that you are interested in $\exp\{0.2\theta_1 - 0.3\theta_2\}$

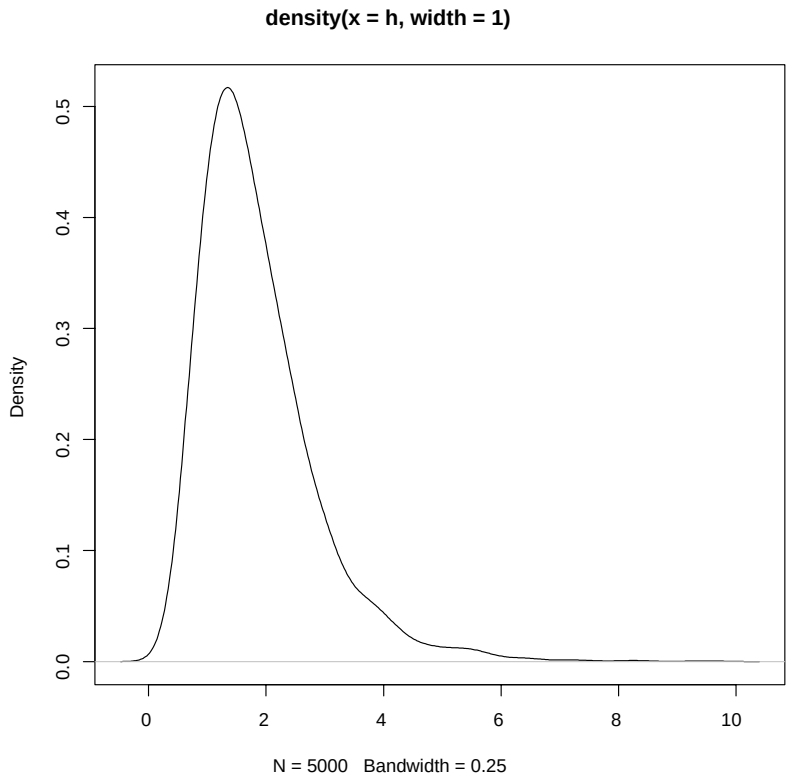
and the posterior density of θ is multivariate normal with mean μ and variance Σ.

> mu <- c(1,-1);Sigma <- matrix(c(2,-0.6,-0.6,1),nrow=2,byrow=T)

> theta <- mvrnorm(5000,mu,Sigma)

```
> h <- rep(0,5000); for(i in 1:5000){h[i] <- exp(.2*theta[i,1]-.3*theta[i,2])}

> hist(h,nclass=50)

> plot(density(h))

> summary(h)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.2722 1.1800 1.6410 1.8580 2.3020 9.6480

> plot(density(h,width=1)))
```

Figure 10: Posterior Density of $\exp\{0.2\theta_1 - 0.3\theta_2\}$



When a distribution can be sampled with a single call, such as mvrnorm, it is called "available". Most posterior distributions are not available. So, what to do?

The answer, since about 1990, is Markov Chain Monte Carlo or MCMC.

Principles of MCMC(pps 192-226)

The state of a markov chain is a random variable indexed by t, say, θ_t. The state distribution is the distribution of θ_t, $p_t(\theta)$.A stationary distribution of the chain is a distribution p such that, if $p_t(\theta) = p$ then $p_{t+s}(\theta) = p$ for all $s > 0$. Under certain conditions a chain will

1. Have a unique stationary distribution.

2. Converge to that stationary distribution as $t \to \infty$. For example, when the sample space for θ is discrete, this means

$$P(\theta_t = j) \to p_j \quad \text{as} \quad t \to \infty.$$

3. Be ergodic. This means that averages of successive realizations of θ will converge to their expectations with respect to p.

A chain is characterized by its transition kernel whose elements provide the conditional probabilities of θ_{t+1} given the values of θ_t. The kernel is denoted by $K(x, y)$.

Example: A 2 State Chain

$$K = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}.$$

When $\theta_t = 1$ then $\theta_{t+1} = 1$ with probability $1-\alpha$ and equals 2 with probability α. For a chain that has a stationary distribution powers of K converge to a constant matrix whose rows are p. For the 2 state chain K^t takes the form

$$K^t = \frac{1}{\alpha+\beta}\begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} + \frac{(1-\alpha-\beta)^t}{\alpha+\beta}\begin{bmatrix} \alpha & -\alpha \\ -\beta & \beta \end{bmatrix}.$$

which converges geometrically fast to a matrix with rows equal to $(\beta/(\alpha+\beta), \alpha/(\alpha+\beta))$.

The stationary distribution of this chain is

$$\Pr(\theta = 1) = \beta/(\alpha+\beta) \tag{30}$$

$$\Pr(\theta = 2) = \alpha/(\alpha+\beta) \tag{31}$$

Example: An Autoregressive Process:

$$K(x,y) = \frac{1}{\sqrt{2\pi}}\exp\{-(1/2)(y-\rho x)^2\}$$

A stationary distribution of the chain, p, satisfies

$$\begin{aligned} p &= pK \\ \text{or} \qquad p(y) &= \int_x K(x,y)p(x)dx. \end{aligned} \tag{*}$$

To check that some $p(.)$ is a stationary distribution of the chain defined by $K(,)$ show it satisfies (*). To prove that $p(y) = n(0, 1-\rho^2)$ is a stationary distribution of the chain with kernel

$$K(x,y) = \frac{1}{\sqrt{2\pi}} e^{-(y-\rho x)^2/2}.$$

Try (*)

$$\begin{aligned} \textstyle\int_x K(x,y)p(x)dx = & \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(y-\rho x)^2/2} \frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}} e^{-(1-\rho^2)x^2/2} dx \\ = & \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-\rho y)^2/2} \frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}} e^{-(1-\rho^2)y^2/2} dx \\ = & \frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}} e^{-(1-\rho^2)y^2/2} = p(y). \end{aligned}$$

The Essence of MCMC

We wish to sample from $p(\theta|y)$. Then let $p(\theta|y)$ be thought of as the stationary distribution of a markov chain and find a chain having this p as its unique stationary distribution. This can be done in many ways!

Then: RUN THE CHAIN until it has converged to p. This means choosing an initial value θ_1 then sampling θ_2 according to the relevant row of K then sampling θ_3 using the relevant row of K .....

When it has converged, realizations of θ have distribution $p(\theta|y)$. They are identically, but not independently, distributed. To study properties of p use the ergodic theorem. e.g.

$$\Sigma_{s=1}^{nrep} I(\theta_{t+s} > 0)/nrep \to P(\theta > 0) \text{ as } nrep \to \infty,$$

where $I(.)$ is the indicator function.

Probability texts focus on the question

> Given a chain finds its stationary distribution(s)

For MCMC the relevant question is

> Given a distribution find a chain that has that distribution as its stationary distribution.

Finding a chain that will do the job.

When θ is scalar this is not an issue – just draw $p(\theta|y)$!

When θ is vector valued with elements $\theta_1, \theta_2, ...\theta_k$ the most intuitive and widely used algorithm for finding a chain with $p(\theta|y)$ as its stationary distribution is the Gibbs Sampler.

p has k univariate component conditionals e.g. when $k = 2$ these are $p(\theta_2|\theta_1)$ and $p(\theta_1|\theta_2)$. A step in the GS samples in turn from the component conditionals. For example, for $k = 2$, the algorithm is

1. choose θ_1^0
2. sample θ_2^1 from $p(\theta_2|\theta_1^0)$
3. sample θ_1^1 from $p(\theta_1|\theta_2^1)$

4 update the superscript by 1 and return to 2.

Steps 2 and 3 described the transition kernel K.

Succesive pairs θ_1, θ_2 are points in the sample space of θ. The successive points tour the sample space. In stationary equilibrium they will visit each region of the space in proportion to its posterior probability.

Next is a graph showing the first few realizations of θ of a Gibbs sampler for the bivariate normal distribution, whose components conditionals are, as is well known, univariate normal.

The second figure has contours of the target (posterior) distribution superimposed.

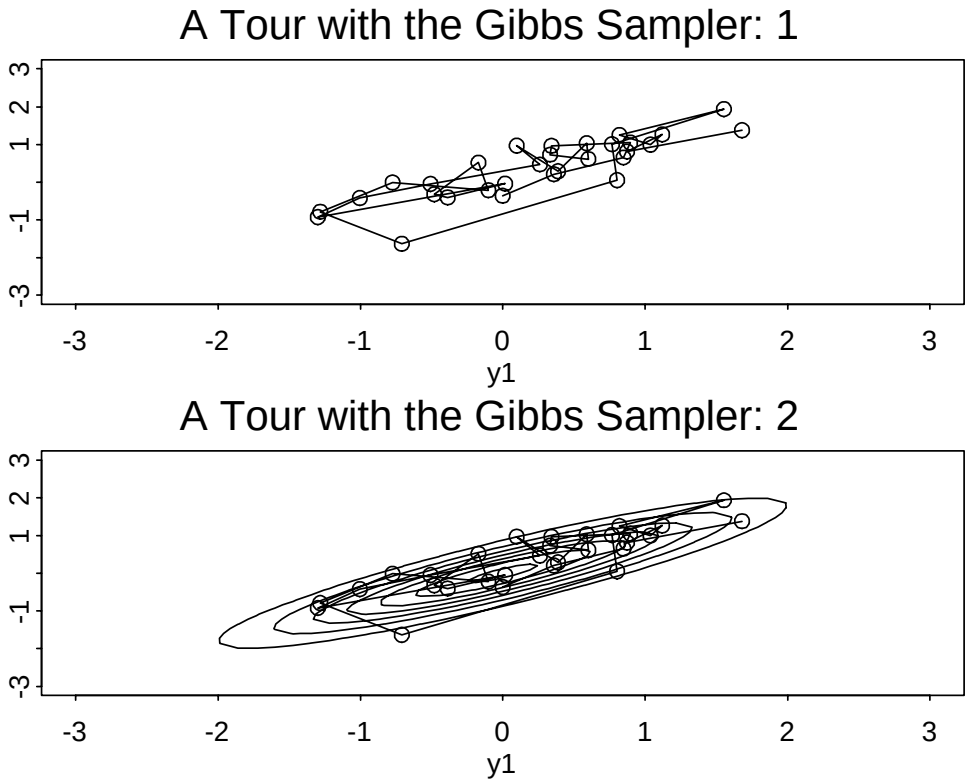

A Tour with the Gibbs Sampler: 1
3
2
1
-1
-3
-3
-2
-1
0
1
2
3
y1
A Tour with the Gibbs Sampler: 2
3
2
1
-1
-3
-3
-2
-1
0
1
2
3
y1

Figure 11:

Gibbs Sampler and Data Augmentation

Data augmentation enlarges the parameter space. Convenient when there is a latent data model.

For example in the probit model

$$y^* = x\beta + \varepsilon, \quad \varepsilon \sim n(0,1) \tag{32}$$

$$y = I_{\{y^*>0\}} \tag{33}$$

Data is y, x. Parameter is β. Enlarge parameter space to β, y^* and consider Gibbs algorithm.

1. $p(\beta|y^*, y) = p(\beta|y^*) = n(b, (X'X)^{-1})$

2. $p(y^*|y, \beta) =$ truncated normals.

Both steps easy.

For another example consider optimal job search. Agents receive job offers and accept the first offer to exceed a reservation wage w^*. The econometrician observes the time to acceptance, t, and the accepted wage, w^a. If offers come from a distribution function $F(w)$ (with $\overline{F} = 1 - F$) and arrive in a Poisson process of rate λ. Duration and accepted wage have joint density

$$\lambda e^{-\lambda \overline{F}(w^*)t} f(w^a); \quad w^a \geq w^*, t \geq 0.$$

This is rather awkward. But consider latent data consisting of the *rejected* wages (if any) and the times at which these offers were received. Let $\theta = (\lambda, w^*)$ plus any parameters of the wage offer distribution and let w, s be the rejected offers and their times of arrival. Data augmentation includes w, s as additional parameters and a Gibbs algorithm would sample in turn from $p(\theta|w, s, w^a, t)$ and $p(w, s|\theta, w^a, t)$ both of which take a very simple form.

A judicious choice of latent data radically simplifies inference about quite complex structural models.

Since about 1993 the main developments have been

- Providing proofs of convergence and ergodicity for broad classes of methods – such as the Gibbs sampler – for finding chains to solve classes of problem.

- Providing effective MCMC algorithms for particular classes of model. In the econometrics journals these include samplers for, e.g. discrete choice models; dynamic general equilibrium models; VARs; stochastic volatility models etc. etc.

But the most important development has been the production of black box general purpose software that enables the user to input his model and data and receive MCMC realizations from the posterior as output without the user worrying about the particular chain that is being used for his problem. (This is somewhat analogous to the development in the frequentist literature of general purpose function minimization routines.)

This development has made MCMC a feasible option for the general applied economist.

Practical MCMC(pps 222-224 and Appendices 2 and 3)

Of the packages available now probably the most widely used is BUGS which is freely distributed from

http://www.mrc-bsu.cam.ac.uk/bugs/

BUGS stands for Bayesian analysis Using the Gibbs Sampler, though in fact it uses a variety of algorithms and not merely GS.

As with any package you need to provide the program with two things:

- The model
- The data

Supplying the data is much as in any econometrics package – you give it the y's and the x's and any other relevant data, for example censoring indicators.

To supply the model you do not simply choose from a menu of models. BUGS is more flexible in that you can give it any model you like!. (Though there are some models that require some thought before they can be written in a way acceptable to BUGS.)

For a Bayesian analysis the model is, of course, the likelihood and the prior.

The model is supplied by creating a file containing statements that closely correspond to the mathematical representation of the model and the prior.

Here is an example of a BUGS model statement for a first order autoregressive model with autoregression coefficient ρ, intercept α and error precision τ.

```
model{
for( i in 2:T){y[i] ~dnorm(mu[i], tau)
mu[i] <- alpha + rho * y[i-1]
}
alpha ~dnorm(0, 0.001)
rho ~dnorm(0, 0.001)
tau ~dgamma(0.001,0.001)
}
```

Lines two and three are the likelihood. Lines five, six and seven are the prior. In this case α, ρ and τ are independent with distributions having low precision (high variance). For example ρ has mean zero and standard deviation $1/\sqrt{0.001} = 32$.

Another BUGS program, this time for an overidentified two equation recursive model.

Model

$$\begin{aligned} y_1 &= b_0 + b_1 y_2 + \varepsilon_1 \\ y_2 &= c_0 + c_1 z_1 + c_2 z_2 + \varepsilon_2. \end{aligned}$$

#2 equation overidentified recursive model with 2 exogenous variables.

```
# Modelled as a restricted reduced form.
model{
for(i in 1:n){
y[i,1:2] ~dmnorm(mu[i,],R[,])
mu[i,1] <- b0 + b1*c0 + b1*c1*z[i,1] + b1*c2*z[i,2]
mu[i,2] <- c0 + c1*z[i,1] + c2*z[i,2]
}
R[1:2,1:2] ~dwish(Omega[,],4)
b0 ~dnorm(0,0.0001)
b1 ~dnorm(0,0.0001)
c0 ~dnorm(0,0.0001)
c1 ~dnorm(0,0.0001)
```

```
c2 ~dnorm(0,0.0001)
}
```

Summary Output Table

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
R[1,1]	0.9647	0.04334	4.708E-4	0.8806	0.9645	1.052	2501	7500
R[1,2]	0.0766	0.03233	3.458E-4	0.01316	0.077	0.1396	2501	7500
R[2,1]	0.0766	0.03233	3.458E-4	0.01316	0.077	0.1396	2501	7500
R[2,2]	1.055	0.04723	5.355E-4	0.9648	1.054	1.15	2501	7500
b0	-0.0136	0.04159	7.906E-4	-0.09936	-0.012	0.064	2501	7500
b1	0.6396	0.2248	0.004277	0.2641	0.6149	1.146	2501	7500

c0 0.0407 0.03111 5.562E-4 -0.01868 0.0400 0.1021 2501 7500

c1 0.1442 0.0284 4.031E-4 0.08972 0.1439 0.2008 2501 7500

c2 0.1214 0.02905 4.623E-4 0.06608 0.1214 0.1797 2501 7500

0.1 An Application of IV Methods to Wages and Education

$$wages = \alpha + \beta.education$$

Education is measured in years and "wages" is the logarithm of the weekly wage rate.

β is the proportionate return to an additional year of education

Rate of return probably between 5 and 30% implying β of the order of 0.05 to 0.30.

BUT education presumably endogenous.

Quarter of birth as instrument as in Angrist and Krueger[1991]. A

> ".....children born in different months of the years start school at different ages, while compulsory schooling laws generally require students to remain in school until their sixteenth or seventeenth birthday. In effect, the interaction of school entry requirements and compulsory schooling laws compel(s) students born in certain months to attend school longer than students born in other months."

The model uses quarter of birth as three instrumental variables. Let q_1, q_2, and q_3 be such that $q_j = 1$ if the agent was born in quarter j and is zero otherwise, and write

$$\text{educ} = \gamma + \delta_1 q_1 + \delta_2 q_2 + \delta_3 q_3 + \varepsilon_1 \tag{34}$$

This model implies that the expected education of someone born in quarter j is $\gamma + \delta_j, j = 1, 2, 3$ So the δ_j are the differences in average education between someone born in quarter j and someone born in the fourth quarter. The second structural equation relates wages to education and we write it as

$$\text{wage} = \alpha + \beta \text{ educ} + \varepsilon_2 \tag{35}$$

since we would expect the relation between education and wages to be monotone and, at least roughly, linear, perhaps. This is an overidentified recursive model. It is overidentified because there are, in fact, three instrumental variables, q_1, q_2 and q_3 but only one right hand endogenous variable, education. Under an assumption of bivariate normality (not necessary) for $\varepsilon_1, \varepsilon_2$ it can be simulated using the BUGS program given earlier.

$n = 35,805$ men born in 1939 (subset of the AK data)

The average number of years of education was 13 and the median number was 12, with 59 men having zero years and 1322 having the maximum education of 20 years. 38% of the men had 12 years, corresponding to completion of high school.

Quarter of birth	Years of education
first	13.003
second	13.013
third	12.989
fourth	13.116

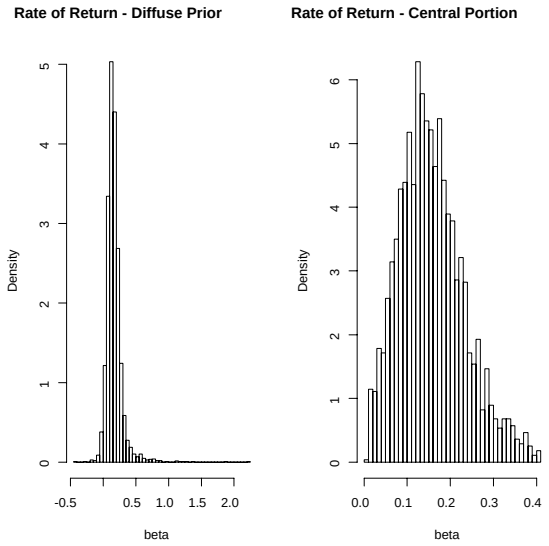
(36)

The differences are small; the difference between 13.116 and 12.989 is about six and half weeks of education and this is about one percent of the median years of education in the whole sample.

Instruments qualify as "weak"

0.1.1 Diffuse Prior.

The sampler was run through 2000 steps with the first 1000 discarded and three parallel chains were run to check convergence – this gives a total of 3000 realizations from the joint posterior. The first figure shows the marginal posterior for β under the diffuse prior setup. The left frame gives the entire histogram for all 3000 realizations and the right frame provides a close-up of the middle 90% of the distribution.



Posterior Distribution of the Rate of Return — Diffuse Prior

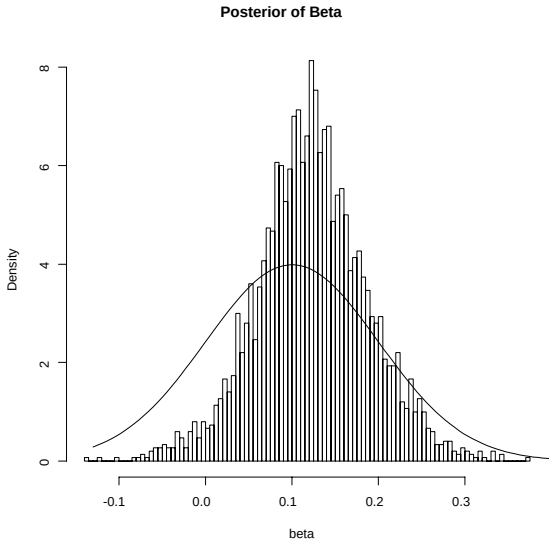
It can be seen that the distribution is very dispersed; the smallest realization (of 3000) corresponds to a return (a loss) of minus 40% and the largest to a gain of 220%. An (approximate) 95% HPD interval runs from a rate of return of essentially zero to one of about 45%.

Distribution not normal –thick tailed. and it is also very far from the prior which, if superimposed on the left hand frame, would look almost flat.

We have learned something but not much. As can be seen on the right frame, the posterior points towards rates of return of the order of 10 to 20% a year, but very considerable uncertainty remains. The ideal might be to be able to pin down the rate of return to within one or two percentage points, but we are very far from this accuracy. The posterior mean of β under this prior is 17% and the median is 15%. 75% of all realizations lie within the range from 10% to 21%.

0.1.2 Informative Prior

For contrast let us see what happens when we use the relatively informative prior in which β is $n(0.10, 100)$ and the three δ coefficients are constrained to be similar. We again ran the sampler for 2000 steps three times and retained the final 1000 of each chain giving 3000 realizations from the joint posterior distribution. The marginal posterior histogram for β is shown in the next figure, with the prior density superimposed as the solid line.



The Posterior Distribution of β with an Informative Prior

0.1.3 Comparison of Posteriors Under Diffuse and Informative Priors.

For a numerical comparison we can look at the 95% HPD intervals which are

	lower	upper
diffuse	0	46%
informative	−1%	25%

The length of the confidence interval has been sharply reduced, but it is still very wide. Another comparison is to look at the quantiles and these, expressed in percentages, are

	min	q_{25}	q_{50}	mean	q_{75}	max
diffuse	−40	10	15	17	21	220
informative	−15	8	12	12	16	37

0.1.4 Conclusions

What do we conclude?

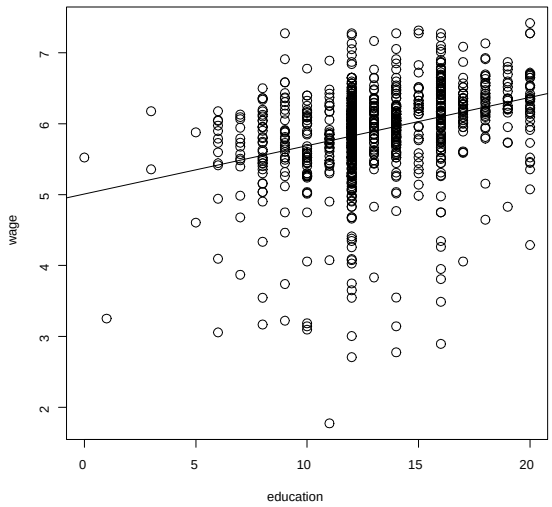
- It appears that even with about $36,000$ observations and a simple model with very few parameters there is not enough information to make precise estimates of the marginal rate of return. This is true even under relatively informative prior beliefs. The data transform such beliefs into less dispersed and more peaked posteriors but these are still not adequate to provide an accurate assessment of the rate of return.

- Another valuable conclusion is that, even with this sample size, it is vital to compute exact posterior distributions that reveal the true nature of our uncertainty.

- A third implication is that by using MCMC methods such exact inferences can be made very easily.

0.2 Is Education Endogenous?

These rather imprecise inferences about a parameter that is of major importance, the rate of return to investment in education, are disappointing. It would have been so much simpler to regress (log) wages on education by least squares. Here is what happens if you do. The next figure plots a randomly selected sub-sample of size 1000 of the education and wage data with the least squares fit (using the whole sample) superimposed upon it.



Education and Wages, with LS Fit

The least squares line is

$$\text{wage} = 5.010(.014) + 0.0680(.001)\text{education}$$

The coefficients are very sharply determined and the posterior mean, median and mode of the rate of return is 6.8% with a standard error of one tenth of one percent. Why can't we accept this estimate? The answer is, of course, that we have been presuming that education is endogenous, that its effect on wages is confounded with that of the numerous other potential determinants of wages. And under this belief the least squares estimates are (very precisely) wrong. But there is still hope for the least squares estimate since we have not, yet, shown that this presumption is true!

Consider how we might check to see whether education really is endogenous. We reproduce the structural form of the model for convenience here. It is

$$\begin{aligned} educ_i &= \gamma + \delta z_i + \varepsilon_{1i} \\ wage_i &= \alpha + \beta\ educ_i + \varepsilon_{2i}, \end{aligned}$$

where z stands for the quarter of birth instruments. We showed earlier that if ε_2 and ε_1 are uncorrelated then the second equation is a regression, the system is fully recursive, and standard regression methods, like least squares, can be used to make inferences about β. But if ε_2 and ε_1 are correlated then these methods are inappropriate and we must proceed as above with the consequent disappointing precision. So why don't we see if these errors are correlated? One way of doing this is to look at the posterior distribution of the correlation coefficient of ε_2 and ε_1. We can do this as follows. For purposes of our calculation we represented the model as the restricted reduced form

$$\begin{aligned} educ_i &= \gamma + \delta z_i + \nu_{1i} \\ wage_i &= \alpha^* + \beta\delta z_i + \nu_{2i}. \end{aligned}$$

The relation between the structural and reduced form errors is

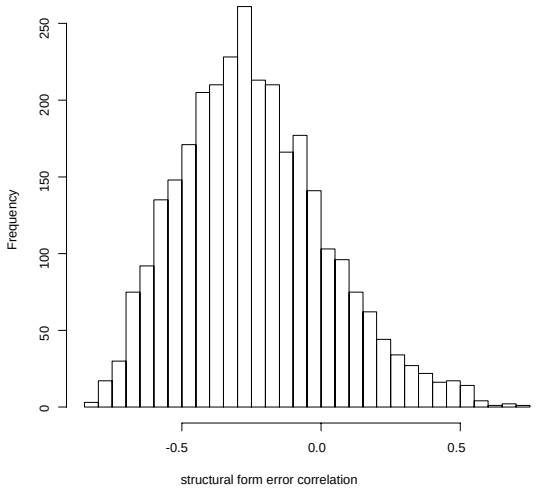
$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \tag{37}$$

or $\varepsilon = B\nu$. Thus $V(\varepsilon) = BV(\nu)B' = BP^{-1}(\nu)B'$, where $P(\nu)$ is the precision of the reduced form errors which, in our BUGS program, we

denoted by q, thus $V(\varepsilon) = Bq^{-1}B'$. We can now use this equation to express the correlation between the structural form errors in terms of β and the reduced form precision matrix q, and we find

$$\rho_{\varepsilon_2,\varepsilon_1} = \frac{-\beta q_{22} - q_{12}}{\sqrt{q_{22}(q_{11} + 2\beta q_{12} + \beta^2 q_{22})}}.$$

Finally, to inspect the posterior distribution of ρ we substitute the 3000 realizations of the four parameters on the right into this formula and the result is the same number of realizations from the posterior distribution of ρ whose histogram is shown below.



Correlation of the Structural Form Errors

The mean and median correlation are both about -0.25 but the standard deviation is 0.26 and 517 out of 3000 realizations are positive. The posterior suggests that the structural form errors are negatively correlated and this is a bit surprising on the hypothesis that a major element of both ε_1 and ε_2 is "ability" and this variable tends to affect positively both education and wages. But the evidence is very far from conclusive.