Introduction to Bayesian Estimation

June 1, 2012

The Plan

1. What is Bayesian statistics? How is it different from frequentist methods?
2. 4 Bayesian Principles
3. Priors: From prior information to prior distributions

The material in today's lecture can be found in any good text book on Bayesian statistics. One good example is Christian P. Robert's *The Bayesian Choice: From decision theoretic foundations to Computational Implementation*

Probability and statistics; What's the difference?

Probability is a branch of mathematics

- There is little disagreement about whether the theorems follow from the axioms

Statistics is an inversion problem: What is a good probabilistic description of the world, given the observed outcomes?

- There is some disagreement about how we interpret data/observations

Why probabilistic models?

Is the world characterized by randomness?

- Is the weather random?
- Is a coin flip random?
- ECB interest rates?

What is the meaning of probability, randomness and uncertainty?

Two main schools of thought:

- The classical (or frequentist) view is that probability corresponds to the frequency of occurrence in repeated experiments
- The Bayesian view is that probabilities are statements about our state of knowledge, i.e. a subjective view.

The difference has implications for how we interpret estimated statistical models and there is no general agreement about which approach is “better”.

Frequentist vs Bayesian statistics

Variance of estimator vs variance of parameter

- Bayesians think of parameters as having distributions while frequentist conduct inference by thinking about the variance of an *estimator*

Frequentist confidence intervals:

- If point estimates are the truth and with repeated draws from the population of equal sample length, what is the interval that $\widehat{\theta}$ lies in 95 per cent of the time?

Bayesian probability intervals:

- Conditional on the observed data, a prior distribution of θ and a functional form (i.e. model), what is the shortest interval that with 95% contains θ?

The main conceptual difference is that Bayesians condition on the data while frequentist design procedures that work well *ex ante.*

Frequentist vs Bayesian statistics

Coin flip example:

After flipping a coin 10 times and counting the number of heads, what is the probability that the next flip will come up heads?

- A Bayesian with a uniform prior would say that the probability of the next flip coming up heads is $x/10$ and where x is the number of times the flip came up heads in the observed sample
- A frequentist cannot answer this question: A frequentist confidence interval can only be used to compute the probability of having observed the sample conditional on a given null hypothesis about the probability of heads vs tails.

Bayesian statistics

"Bayesianism has obviously come a long way. It used to be that could tell a Bayesian by his tendency to hold meetings in isolated parts of Spain and his obsession with coherence, self-interrogations, and other manifestations of paranoia. Things have changed..."

Peter Clifford, 1993

Quote arrived here via Jesus Fernandez-Villaverde (U Penn)

Bayesian statistics

Bayesians used to be fringe types, but are now more or less the mainstream in macro,

- This is largely due to increased computing power

Bayesian methods have several advantages:

- Facilitates incorporating information from outside of sample
- Easy to compute confidence/probabilty intervals of functions of parameters
- Good small sample properties

The subjective view of probability

What does subjective mean?

- Coin flip: What is prob(heads) after flip but before looking?

Important: Probabilities can be viewed as statements about our knowledge of a parameter, even if the parameter is a constant

- E.g. treating risk aversion as a random variable does not imply that we think of it as varying across time.

The subjective view of probability

Is a subjective view of probability useful?

- Yes, it allow people to incorporate information from outside the sample
- Yes, because subjectivity is always there (it just more hidden in frequentist methods)
- Yes, because one can use “objective”, or non-informative priors
- And: Sensitivity to priors can be checked

4 Principle of Bayesian Statistics

1. The Likelihood Principle
2. The Sufficiency Principle
3. The Conditionality Principle
4. The Stopping Rule Principle

These are principles that appeal to “common sense”.

The Likelihood Principle

All the information about a parameter from an experiment is contained in the likelihood function

The Likelihood Principle

The likelihood principle can be derived from two other principles:

- The Sufficiency Principle
  - If two samples imply the same sufficient statistic (given a model that admits such a thing), then they should lead to the same inference about θ
  - Example: Two samples from a normal distribution with the same mean and variance should lead to the same inference about the parameters μ and σ^2
- The Conditionality Principle
  - If there are two possible experiments on the parameter θ is available, and one of the experiments is chosen with probability p then the inference on θ should only depend on the chosen experiment

The Stopping Rule Principle

The Stopping Rule Principle is an implication of the Likelihood Principle

If a sequence of experiments, $\varepsilon 1, \varepsilon 2, ...$, is directed by a stopping rule, τ, which indicates when the experiment should stop, inference about θ should depend on τ only through the resulting sample.

The Stopping Rule Principle: Example

In an experiment to find out the fraction of the population that watched a TV show, a researcher found 9 viewers and and 3 non-viewers.

Two possible models:

1. The researcher interviewed 12 people and thus observed $x \sim B(12, \theta)$
2. The researcher questions N people until he obtained 3 non-viewers with $N \sim Neg(3, 1 - \theta)$

The likelihood principle implies that the inference about θ should be identical if either model was used. A frequentist would draw different conclusions depending on wether he thought (1) or (2) generated the experiment.
In fact, $H^0 : \theta > 0.5$ would be rejected if experiment is (2) but not if it is (1).

The Stopping Rule Principle

"...a hypothesis which may be true may be rejected because it has not predicted observable results which have not occurred."
Jeffreys (1961)

"I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right"
Savage (1962)

Quotes (again) arrived here via Jesus Fernandez-Villaverde (U Penn)

Main concepts and notation

The main components in Bayesian inference are:

- Data (observables) $Z^T \in \mathbb{R}^{T \times n}$
- A model:
  - Parameters $\theta \in \mathbb{R}^k$
  - A prior distribution $p(\theta) : \mathbb{R}^k \longrightarrow \mathbb{R}^+$
  - Likelihood function $p(Z \mid \theta) : \mathbb{R}^{T \times n} \times \mathbb{R}^k \longrightarrow \mathbb{R}^+$

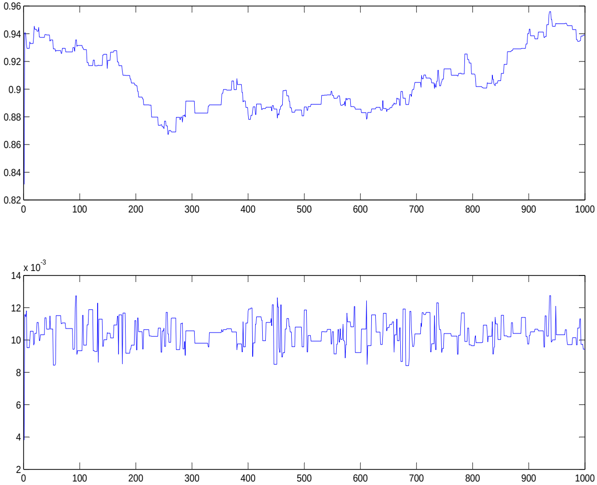
The end product of Bayesian statistics

Most of Bayesian econometrics consists of simulating distributions of parameters using numerical methods.

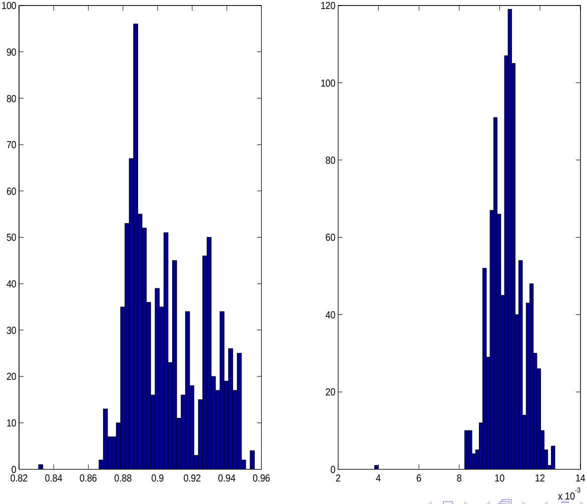
- A simulated posterior is a numerical approximation to the distribution $p(Z \mid \theta)p(\theta)$
- This is useful since the the distribution $p(Z \mid \theta)p(\theta)$ (by Bayes' rule) is proportional to $p(\theta \mid Z)$
- We rely on ergodicity, i.e. that the moments of the constructed sample correspond to the moments of the distribution $p(\theta \mid Z)$

The most popular procedure to simulate the posterior is called the Random Walk Metropolis Algorithm

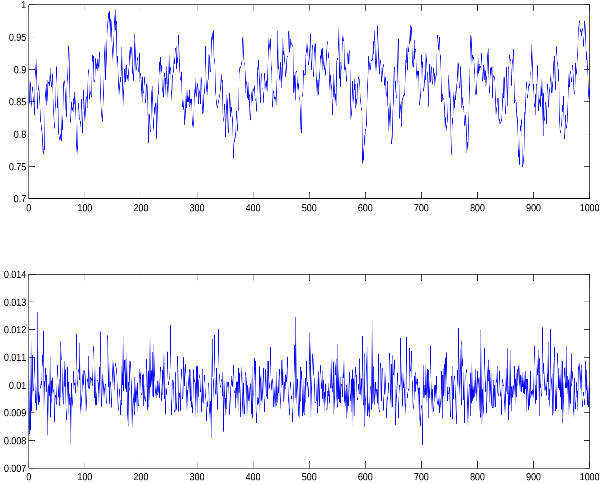
The MCMC with J=1000



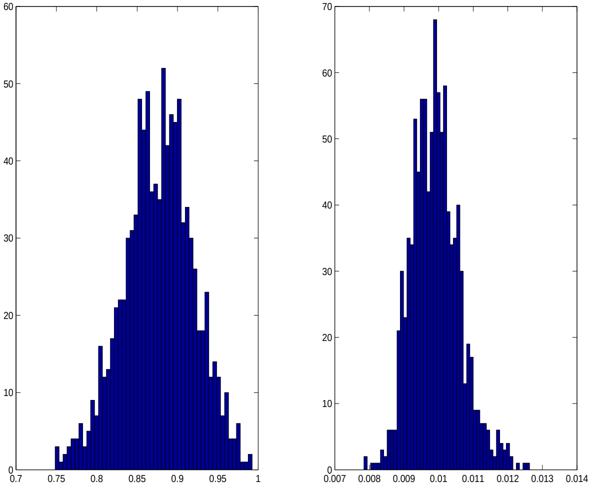
The Histograms of MCMC with J=1000



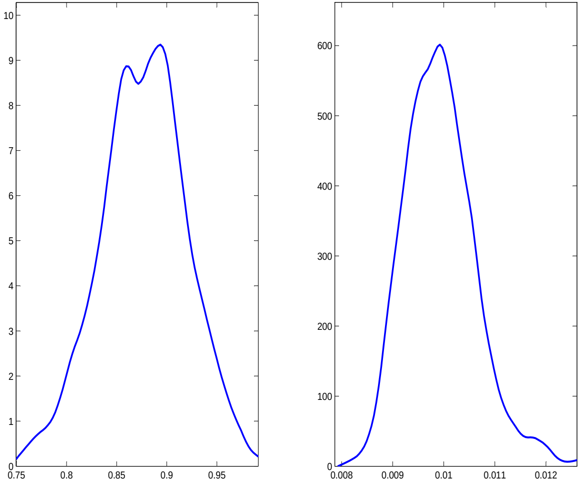
The MCMC with J=1000 000



The Histograms of MCMC with J=1000 000



Estimated posterior density with J=1000



Choosing priors

How influential the priors will be for the posterior is a choice:

- It is always possible to choose priors such that a given result is achieved no matter what the sample information is (i.e. dogmatic priors)
- It is also possible to choose priors such that they do not influence the posterior (i.e. so-called non-informative priors)

Important: Priors are a choice and **must** be motivated.

Combining prior and sample information

Sometimes we know more about the parameters than what the data tells us, i.e. we have some prior information.

- For a DSGE model, we may have information about "deep" parameters
  - Range of some parameters may be restricted by theory, e.g. risk aversion should be positive
  - Discount rate is inverse of average real interest rates
  - Price stickiness can be measured by surveys
- We may know something about the mean of a process

How do we combine prior and sample information?

Bayes' theorem:

$$
\begin{aligned}
P(\theta \mid Z)\, P(Z) &= P(Z \mid \theta)\, P(\theta) \\
&\Leftrightarrow \\
P(\theta \mid Z) &= \frac{P(Z \mid \theta)\, P(\theta)}{P(Z)}
\end{aligned}
$$

- Since $P(Z)$ is constant (conditional on a particular model), we can use $P(Z \mid \theta)\, P(\theta)$ as the posterior likelihood (a likelihood function is any function that is proportional to the probability).

We now need to choose $P(\theta)$

Choosing prior distributions

The beta distribution is a good choice when parameter is in [0,1]

$$P(x) = \frac{(1-x)^{b-1}\, x^{a-1}}{B(a,b)}$$

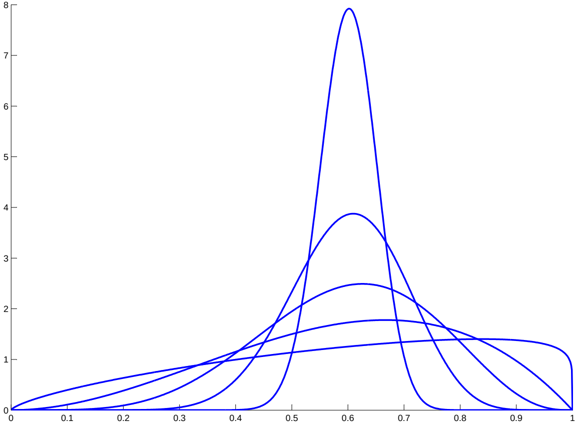
where

$$B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$$

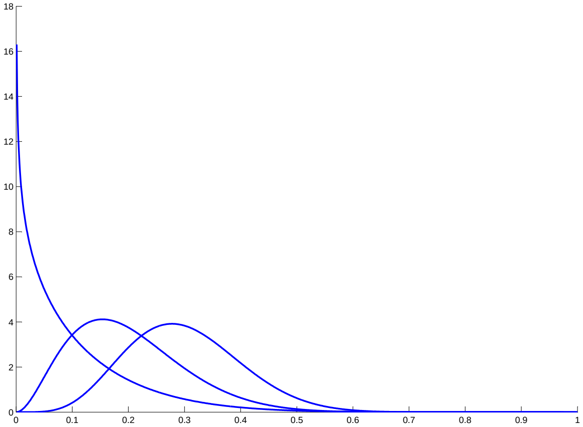
Easier to parameterize using expression for mean, mode and variance:

$$\begin{aligned}\mu &= \frac{a}{a+b}, \quad \widehat{x} = \frac{a-1}{a+b-2} \\ \sigma^2 &= \frac{ab}{(a+b)^2\,(a+b+1)}\end{aligned}$$

Examples of beta distributions holding mean fixed



Examples of beta distributions holding s.d. fixed



Choosing prior distributions

The inverse gamma distribution is a good choice when parameter is positive

$$P(x) = \frac{b^a}{\Gamma(a)}(1/x)^{a+1}\exp(-b/x)$$

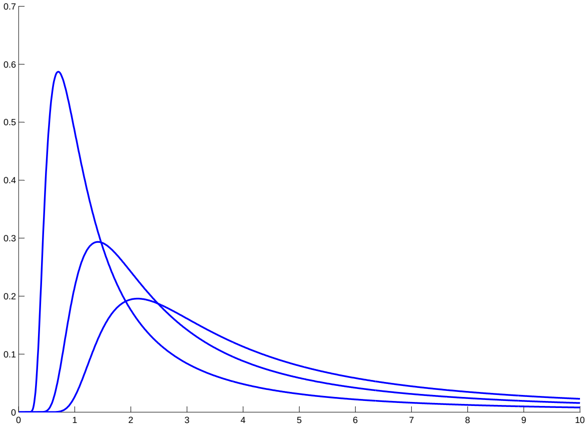
where

$$\Gamma(a) = (a-1)!$$

Again, easier to parameterize using expression for mean, mode and variance:

$$\begin{aligned} \mu &= \frac{b}{a-1}; a > 1, \quad \widehat{x} = \frac{b}{a+1} \\ \sigma^2 &= \frac{b^2}{(a-1)^2\,(a-2)}; a > 2 \end{aligned}$$

Examples of inverse gamma distributions



Conjugate Priors

Conjugate priors are a particularly convenient:

- Combining distributions that are members of a conjugate family result in a new distribution that is a member of the same family

Useful, but only so far as that the priors are chosen to actually reflect prior beliefs rather than just for analytical convenience

Conjugate Priors: Examples

$p(x\|\theta)$	$p(\theta)$	$p(\theta\|x)$
$N(\theta, \sigma^2)$	$N(\mu, \tau^2)$	$N(\left(\sigma^2 + \tau^2\right)^{-1}\left(\sigma^2\mu + \tau^2 x\right), \left(\sigma^2 + \tau^2\right)^{-1}\sigma^2\tau^2)$
$P(\theta)$	$G(\alpha, \beta)$	$G(\alpha + x, \beta + 1)$
$G(\nu, \theta)$	$G(\alpha, \beta)$	$G(\alpha + \nu, \beta + x)$

Improper Priors

Improper priors are priors that are not probability density functions in the sense that they do not integrate to 1.

- Can still be used as a form of uninformative priors for some types of analysis
- The uniform distribution $U(-\infty, \infty)$ is popular
- Mode of posterior then coincide with MLE.

Summing up

- It is important to understand what the subjective view of probability does and does not imply
- The Bayesian Principles
- Bayesians condition on the data
- Priors are chosen by the researcher and must be motivated

That’s it for today.