SCOTTISH GRADUATE PROGRAMME IN ECONOMICS

ECNM11060

Bayesian Econometrics

An Overview of Bayesian Econometrics

Core concepts in Bayesian econometrics

The next few slides provide a brief overview of:

⇒ Bayes' Theorem

⇒ Bayesian learning and updating

⇒ Prior and posterior distributions

⇒ Predictive inference

⇒ Model comparison and selection

Bayesian Theory

⇒ Begin with general concepts in Bayesian theory before getting to specific models

⇒ If you know these general concepts you will never get lost

⇒ What does an econometrician do?

1. Estimate parameters in a model (e.g., regression coefficients)
2. Compare different models (e.g., hypothesis testing)
3. Predictions

⇒ Bayesian econometrics does these based on a few simple rules of probability

Some important probabilities

- ⇒ Let A and B be two events, $p(B|A)$ is the conditional probability of $B|A$: "summarizes what is known about B given A"
- ⇒ Bayesians use this rule with A = something known or assumed (e.g., the data), B is something unknown (e.g., regression coefficients in a model)
- ⇒ Let $\boldsymbol{y} = (y_1, \ldots, y_N)'$ be observed data, y^* be an unobserved data point (i.e., to be predicted), M_i (for $i = 1, \ldots, m$) be a set of models each of which depends on some parameters, $\theta^{(i)}$
- ⇒ Learning about parameters in a model is based on the posterior density: $p(\theta^{(i)}|M_i, \boldsymbol{y})$
- ⇒ Model comparison based on posterior model probability: $p(M_i|\boldsymbol{y})$
- ⇒ Prediction based on the predictive density $p(y^*|\boldsymbol{y})$

Bayes Theorem

I expect you know basics of probability theory from previous studies, see Appendix B of main textbook if you do not

Definition: Conditional probability

The conditional probability of A given B, denoted by $\Pr(A|B)$, is the probability of event A occurring given event B has occurred

Theorem: Rules of conditional probability including Bayes' Theorem

Let A and B denote two events, then

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

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

Bayes Theorem

Bayes' Theorem

These two rules can be combined to yield **Bayes' Theorem**:

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

Note: Above is expressed in terms of two events, A and B. However, can be interpreted as holding for random variables, A and B with probability density functions replacing the $\Pr()$s in previous formulae.

Learning about parameters in a given model (estimation)

- $\Rightarrow$ Assume a single model which depends on parameters θ
- $\Rightarrow$ Want to figure out properties of the posterior $p(\theta|\boldsymbol{y})$
- $\Rightarrow$ It is convenient to use Bayes' rule to write the posterior in a different way
- $\Rightarrow$ Bayes' rule lies at the heart of Bayesian econometrics:

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

- $\Rightarrow$ Replace B by θ and A by $\boldsymbol{y}$ to obtain:

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

Let's have a closer look...

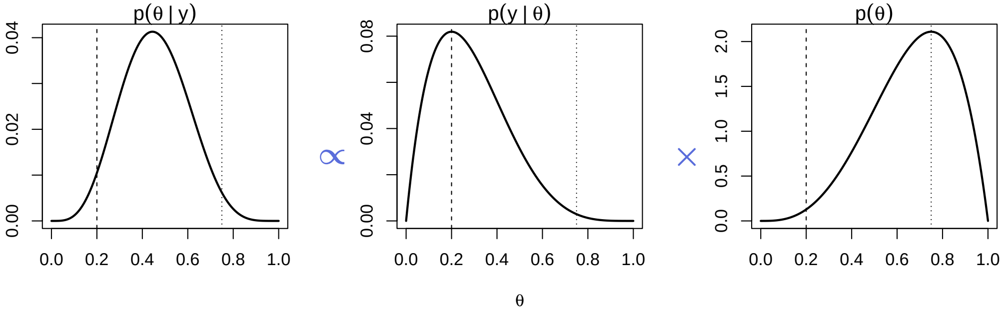
- $\Rightarrow$ Bayesians treat $p(\theta|\boldsymbol{y})$ as being of fundamental interest: "Given the data, what do we know about θ?"
- $\Rightarrow$ Treatment of θ as a random variable is controversial among some econometricians
- $\Rightarrow$ Competitor to Bayesian econometrics, called **frequentist econometrics**, says that θ is not a random variable
- $\Rightarrow$ For estimation can ignore the term $p(\boldsymbol{y})$ since it does not involve θ (basically ensures the density integrates to one)

Posterior is proportional to likelihood times prior

$$\underbrace{p(\theta|\boldsymbol{y})}_{\text{Posterior}} \propto \underbrace{p(\boldsymbol{y}|\theta)}_{\text{Likelihood}} \times \underbrace{p(\theta)}_{\text{Prior}}$$

Bayes Theorem in action

We combine data (likelihood) and non-data (prior) information to obtain an updated (proportional) posterior distribution



Understanding the prior

⇒ $p(\theta)$ does not depend on the data; it contains any non-data information available about θ

⇒ Prior information is controversial aspect since it sounds unscientific

⇒ Bayesian answers (to be elaborated on later):

1. Often we do have prior information and, if so, we should include it (more information is good)
2. Can work with so-called “noninformative” priors
3. Can use hierarchical priors which treat prior hyperparameters as parameters and estimates them
4. Training sample priors or priors obtained from a theoretical model (e.g., DSGE model)
5. Bayesian estimators often have better frequentist properties than frequentist estimators
6. Prior sensitivity analysis

Prediction in a single model

⇒ Prediction based on the **predictive density** $p(y^*|\boldsymbol{y})$

⇒ Since a marginal density can be obtained from a joint density through integration:

$$p(y^*|\boldsymbol{y}) = \int p(y^*, \theta|\boldsymbol{y})d\theta$$

⇒ Term inside integral can be rewritten as:

$$p(y^*|\boldsymbol{y}) = \int p(y^*|\boldsymbol{y}, \theta)p(\theta|\boldsymbol{y})d\theta$$

⇒ Prediction involves the posterior and $p(y^*|\boldsymbol{y}, \theta)$ (more description provided later)

Model comparison (hypothesis testing)

⇒ Models denoted by M_i for $i = 1, \ldots, m$

⇒ M_i depends on parameters $\theta^{(i)}$

Posterior model probability: $p(M_i|y)$

Using Bayes rule with $B = M_i$ and $A = \boldsymbol{y}$ we obtain:

$$p(M_i|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M_i)p(M_i)}{p(\boldsymbol{y})}$$

⇒ $p(M_i)$ is referred to as the **prior model probability**

⇒ $p(\boldsymbol{y}|M_i)$ is called the **marginal likelihood**

Marginal likelihood

⇒ How is marginal likelihood calculated?

⇒ Posterior can be written as:

$$p(\theta^{(i)}|\boldsymbol{y}, M_i) = \frac{p(\boldsymbol{y}|\theta^{(i)}, M_i)p(\theta^{(i)}|M_i)}{p(\boldsymbol{y}|M_i)}$$

⇒ Integrate both sides with respect to $\theta^{(i)}$, use fact that $\int p(\theta^{(i)}|\boldsymbol{y}, M_i)d\theta^{(i)} = 1$ and rearrange:

$$p(\boldsymbol{y}|M_i) = \int p(\boldsymbol{y}|\theta^{(i)}, M_i)p(\theta^{(i)}|M_i)d\theta^{(i)}$$

⇒ Note: marginal likelihood depends only on the prior and likelihood

Posterior odds ratio

⇒ Posterior odds ratio compares two models:

$$PO_{ij} = \frac{p(M_i|\boldsymbol{y})}{p(M_j|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|M_i)p\,(M_i)}{p(\boldsymbol{y}|M_j)p(M_j)}$$

⇒ Note: $p(y)$ is common to both models, no need to calculate

Bayes factors

⇒ Can use fact that $p(M_1|\boldsymbol{y}) + p(M_2|\boldsymbol{y}) + \cdots + p(M_m|\boldsymbol{y}) = 1$ and PO_{ij} to calculate the posterior model probabilities

⇒ For example, suppose $m = 2$ models and you know:

$$p(M_1|\boldsymbol{y}) + p(M_2|\boldsymbol{y}) = 1$$

$$PO_{12} = \frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})}$$

implies

$$p(M_1|\boldsymbol{y}) = \frac{PO_{12}}{1 + PO_{12}}$$

$$p(M_2|\boldsymbol{y}) = 1 - p(M_1|\boldsymbol{y})$$

⇒ The **Bayes factor** is:

$$BF_{ij} = \frac{p(\boldsymbol{y}|M_i)}{p(\boldsymbol{y}|M_j)}$$

Example with Bernoulli distributed data

Example: Deriving posterior when data has Bernoulli distribution

- $\Rightarrow$ Experiment repeated N times
- $\Rightarrow$ Each time the outcome can be "success" or "failure"(e.g., a coin flip)
- $\Rightarrow$ y_j, for $j = 1, .., N$, are random variables for each repetition of experiment
- $\Rightarrow$ Realization of y_j can be 1 (= success) or 0 (= failure)
- $\Rightarrow$ Probability of success is θ (hence probability of failure is $1 - \theta$)
- $\Rightarrow$ The goal is to estimate θ

Example (cont.): The Bernoulli likelihood function

⇒ Notation for things above is: $y_j \in \{0, 1\}$, $0 \leq \theta \leq 1$, and

$$p(y_j|\theta) = \begin{cases} \theta & \text{if } y_j = 1 \\ 1 - \theta & \text{if } y_j = 0 \end{cases}$$

⇒ Let S_N be the number of successes in N repetitions of experiment

⇒ Likelihood function is:

$$\begin{aligned} p(\boldsymbol{y}|\theta) &= \prod_{j=1}^{N} p(y_j|\theta) \\ &= \theta^{S_N}(1-\theta)^{N-S_N} \end{aligned}$$

Example (cont.): The (conjugate) Beta prior

- ⇒ View this likelihood in terms of θ: proportional to pdf of a Beta distribution
- ⇒ A beta-distributed random variable $x \sim \mathcal{B}(a, b)$ has pdf

$$f_{\mathcal{B}}(x; a, b) \propto x^{a-1}(1-x)^{b-1} \text{ with } 0 < x < 1$$

- ⇒ Most common distribution for random variables bounded to lie in $[0, 1]$
- ⇒ Commonly used for parameters which are probabilities (like θ)
- ⇒ Bayesians need prior: Let's use also Beta distribution for prior
- ⇒ Prior beliefs concerning θ are represented by

$$p(\theta) \propto \theta^{a_0-1}(1-\theta)^{b_0-1}$$

Example (cont.): Eliciting a prior

- $\Rightarrow$ The researcher chooses prior hyperparameters $a_0 > 0$ and $b_0 > 0$ to reflect beliefs
- $\Rightarrow$ Properties of Beta distribution imply prior mean is

$$\mathbb{E}(\theta) = \frac{a_0}{a_0 + b_0}$$

- $\Rightarrow$ Suppose you believe, a priori, that success and failure are equally likely
- $\Rightarrow$ $\mathbb{E}(\theta) = \frac{1}{2}$ obtained by setting $a_0 = b_0$
- $\Rightarrow$ If I look on Wikipedia I see $a_0 = b_0 = 2$ has mean at $\mathbb{E}(\theta) = \frac{1}{2}$ but spreads probability widely over interval $[0, 1]$
- $\Rightarrow$ So I might be “relatively noninformative”and choose this for my prior

Example (cont.): A noninformative prior

- $\Rightarrow$ A common choice is $a_0 = b_0 = 0.01$, which yields a very diffuse (weakly informative) prior
- $\Rightarrow$ Special case: $a_0 = b_0 = 1$
  - $\rightarrow$ Then $p(\theta) \propto 1$
  - $\rightarrow$ This corresponds to a $\mathcal{B}(1, 1)$ distribution
  - $\rightarrow$ A uniform distribution on $[0, 1]$
  - $\rightarrow$ All values of θ receive equal prior weight
- $\Rightarrow$ Question: Is a uniform prior truly "noninformative"?
- $\Rightarrow$ See discussion in Zhu et al. (2004)

Example (cont.): Deriving the posterior

$\Rightarrow$ To get posterior multiply likelihood times prior

$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\theta)p(\theta) \\ &\propto \underbrace{\theta^{S_N}(1-\theta)^{N-S_N}}_{\text{Likelihood}} \underbrace{\theta^{a_0-1}(1-\theta)^{b_0-1}}_{\text{Prior}} \\ &\propto \theta^{a_0+S_N-1}(1-\theta)^{b_0+N-S_N-1} \end{aligned}$$

Example (cont.): Interpretation and terminology

$\Rightarrow$ Posterior has same Beta form as prior (terminology = conjugate)

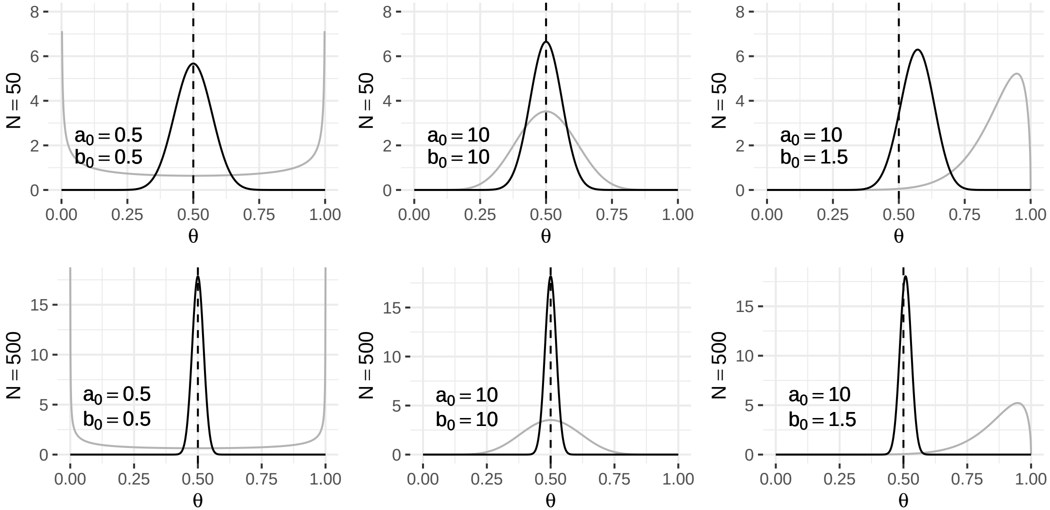
$$\theta|\boldsymbol{y} \sim \mathcal{B}\left(\underbrace{a_0 + S_N}_{a_1}, \underbrace{b_0 + N - S_N}_{b_1}\right)$$

$\Rightarrow$ Posterior has arguments a_1 and b_1 instead of a_0 and b_0

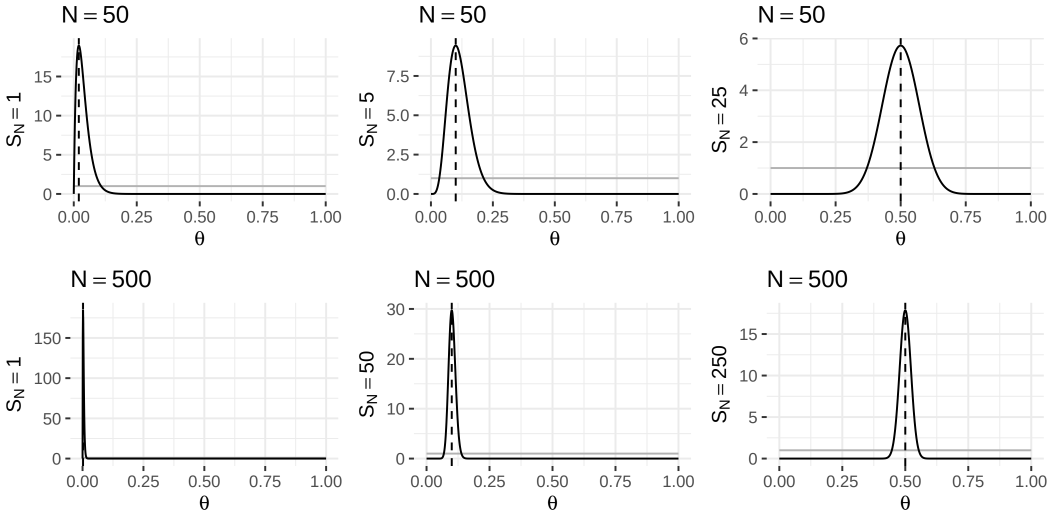
$\Rightarrow$ Arguments have been updated:

- $\rightarrow$ Begin with prior belief (a_0 and b_0) update with data information (S_N and $N - S_N$)
- $\rightarrow$ Posterior combines prior and data information
- $\rightarrow$ "Bayesian learning" = learn about θ by combining prior and data information

Example (cont.): Priors and posteriors



Example (cont.): Priors and posteriors



Example (cont.): Predictive density

- $\Rightarrow$ Derivations of marginal likelihood and predictive density are a bit messier
- $\Rightarrow$ Exercise 7.1 in Bayesian Econometric Methods shows predictive density has Beta-Binomial distribution and shows:

$$\mathbb{E}\left(y^{*}|\boldsymbol{y}\right)=\frac{a_1}{a_1+b_1}$$

- $\Rightarrow$ How do we interpret this?
- $\Rightarrow$ Question: If we run the experiment again what is the probability of getting a success?
- $\Rightarrow$ Answer: $\frac{a_1}{a_1+b_1}$

Summary

- $\Rightarrow$ These few slides have outlined all the basic theoretical concepts required for the Bayesian to learn about parameters, compare models and predict
- $\Rightarrow$ This is an enormous advantage: Once you accept that unknown things (i.e., θ, M_i and y^*) are random variables, the rest of Bayesian approach is non-controversial
- $\Rightarrow$ What are going to do in rest of this course?
- $\Rightarrow$ See how these concepts work in some models of interest
- $\Rightarrow$ First the regression model
- $\Rightarrow$ Then time series models of interest for macroeconomics
- $\Rightarrow$ Bayesian computation

A few final remarks on Bayesian computation

Bayesian Computation

- $\Rightarrow$ How do you present results from a Bayesian empirical analysis?
- $\Rightarrow$ $p(\theta|\boldsymbol{y})$ is a pdf; especially if θ is a vector of many parameters cannot present a graph of it
- $\Rightarrow$ Want features analogous to frequentist point estimates and confidence intervals
- $\Rightarrow$ A common point estimate is the mean of the posterior density (or **posterior mean**)
- $\Rightarrow$ Let $\boldsymbol{\theta}$ be a vector with K elements, $\theta = (\theta_1, .., \theta_K)'$

Posterior expectation

⇒ The posterior mean of any element of $\boldsymbol{\theta}$ is:

$$\mathbb{E}(\theta_i|\boldsymbol{y}) = \int \theta_i p(\theta|y) d\theta$$

Definition B.8: Expected value

⇒ Let $g()$ be a function, then the *expected value* of $g(X)$, denoted $\mathbb{E}[g(X)]$, is defined by:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, p(x)\, dx$$

if X is a continuous random variable (provided $\mathbb{E}[g(X)] < \infty$)

Posterior variance

$\Rightarrow$ Common measure of dispersion is the **posterior standard deviation** (square root of **posterior variance**)

$\Rightarrow$ Posterior variance:

$$\mathbb{V}(\theta_i|\boldsymbol{y}) = \mathbb{E}(\theta_i^2|\boldsymbol{y}) - [\mathbb{E}(\theta_i|y)]^2$$

$\Rightarrow$ This requires calculating another expected value:

$$\mathbb{E}(\theta_i^2|\boldsymbol{y}) = \int \theta_i^2 p(\theta|\boldsymbol{y}) d\theta$$

$\Rightarrow$ Many other possible features of interest...

Important posterior features

- ⇒ All of these posterior features have the form (see Appendix B of main book for definitions)

$$\mathbb{E}\left[g\left(\boldsymbol{\theta}\right)|\boldsymbol{y}\right] = \int g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

  where $g(\boldsymbol{\theta})$ is a **function of interest**

- ⇒ All these features have integrals in them

- ⇒ Marginal likelihood and predictive density also involved integrals

- ⇒ Apart from a few simple cases, it is not possible to evaluate these integrals analytically, and we must turn to the computer

Posterior Simulation

- ⇒ The integrals involved in Bayesian analysis are usually evaluated using simulation methods
- ⇒ Will use several methods later on; here we provide some intuition
- ⇒ Frequentist asymptotic theory uses Laws of Large Numbers (LLN) and a Central Limit Theorems (CLT)
- ⇒ A typical LLN: "consider a random sample, $Y_1, ..Y_N$, as N goes to infinity, the average converges to its expectation" (e.g. $\overline{Y} \to \mu$)
- ⇒ Bayesians use LLN: "consider a random sample from the posterior, $\theta^{(1)}, ..\theta^{(S)}$, as S goes to infinity, the average of these converges to $E[\theta|y]$"
- ⇒ Note: Bayesians use asymptotic theory, but asymptotic in S (under control of researcher) not N

Example: Monte Carlo integration

$\Rightarrow$ Let $\theta^{(s)}$ for $s = 1, .., S$ be a random sample from $p(\theta|\boldsymbol{y})$ and define

$$\widehat{g}_S = \frac{1}{S} \sum_{s=1}^{S} g\left(\theta^{(s)}\right),$$

then $\widehat{g}_S$ converges to $\mathbb{E}\left[g\left(\theta\right)|\boldsymbol{y}\right]$ as S goes to infinity

$\Rightarrow$ **Computational trade-offs**

- $\rightarrow$ Monte Carlo integration approximates $\mathbb{E}\left[g\left(\theta\right)|\boldsymbol{y}\right]$, but only if S were infinite would the approximation error be zero
- $\rightarrow$ We can choose any value for S but larger values of S will increase computational burden

$\Rightarrow$ To gauge size of approximation error, use a CLT to obtain numerical standard error

Statistical software used by Bayesian

- ⇒ Most Bayesians write own codes (e.g., using Matlab, Julia, Python, R or C++) to do posterior simulation
- ⇒ Bayesian work cannot (easily) be done in standard econometric packages like Stata
- ⇒ Stata has some Bayes, but limited (and little for macroeconomics)
- ⇒ Researchers typically post their code on their webpages or GitHub
  - → See my GitHub page here (R) and Ping's webpage here (Matlab)
  - → Colleagues and co-authors who provide excellent code packages: Joshua Chan (Matlab), Florian Huber (R), Karin Klieber (R), Gary Koop (Matlab), Dimitris Korobilis (Matlab), Haroon Mumtaz (Matlab), Michael Pfarrhofer (R), Aubrey Poon (Matlab), Tomasz Woźniak (R)

Learning outside of lectures

- ⇒ Go through the textbook and readings provided
- ⇒ Computational methods are the most important thing for the aspiring Bayesian econometrician to learn
- ⇒ Thus, we devote all of the tutorial hours in this course to computer sessions
- ⇒ Four computer sessions based on four question sheets
- ⇒ Computer code will be provided which will “answer” the questions
- ⇒ Work through/adapt/extend the code
- ⇒ Idea is to develop skills so as to produce your own code or adapt someone else’s for your purposes

Learning outside of lectures

- ⇒ What about proofs/derivations of theoretical results?
- ⇒ In lectures (with a few exceptions) will not do proofs
- ⇒ For example, just state a particular posterior in Normal with formula given for mean and variance
- ⇒ To use Bayesian methods in practice, this is usually all that is needed
- ⇒ But if you want to derive posterior for new model or obtain deeper understanding need to learn necessary tools
- ⇒ These tools best learned by practicing on your own
- ⇒ I will provide Problem Sheets which give practice problems and ask for derivations of some key results
- ⇒ Answers are provided, so I will not formally take them up in lectures or tutorials
- ⇒ Bayesian Econometrics Methods by Chan, Koop, Poirier and Tobias has many more practice problems (and answers)