Bayesian Econometrics

Lecture 1: Introduction to Bayesian Econometrics

Prof. Dr. Kai Carstensen

Kiel University

Winter Term 2024/25

Outline of this lecture

1. Basic Bayesian Theory
2. Natural Conjugate Priors
3. Model Comparison
4. Prediction
5. Computation

This lecture follows Chapter 1 of Koop (2003). Please read this chapter carefully.

References

Strongly recommended:

Gary Koop (2003) Bayesian Econometrics, Wiley

This is the main reference for this and the following chapters. We will closely follow the notation of this textbook. In addition, there is a companion website to the textbook that contains data and Matlab programs.

Valuable:

G. Koop, D.J. Poirier, J.L. Tobias (2007) Bayesian Econometric Methods, Cambridge University Press

Many solved exercises. This book nicely accompanies the tutorial. Matlab programs can be found here.

1. Basic Bayesian Theory

Probabilities

Random variables: A and B.

If continuous, they have probability density function (pdf): $p(A)$ and $p(B)$.

If discrete, they have probability mass function (pmf): $p(A)$ and $p(B)$.

In the following, we will either be general and understand that $p(\cdot)$ is either a pdf or pmf and simply call it probability (even though this is a misuse of language for continuous random variables). Or we discuss the case of continuous random variables and assume you all know that equivalent statements can be made for discrete random variables.

Bayes' rule

Conditional-marginal factorization of joint probability:

$$p(A, B) = p(A|B)p(B)$$

and

$$p(A, B) = p(B|A)p(A)$$

Equating:

$$p(B|A)p(A) = p(A|B)p(B)$$

Solving for $p(B|A)$ yields Bayes' Rule:

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

Bayes' rule: data and parameters

The econometrician has data y and wants to learn about parameter(s) θ.

Since the parameter is unknown, Bayesians treat it as a random variable.

Replace $A \to y$ and $B \to \theta$. Then Bayes' rule becomes

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

Interest rests on $p(\theta|y)$ which answers, in a statistical sense, the question: "Given the data, what do we know about θ?"

Prior, posterior, likelihood

As the marginal density $p(y)$ does not depend on θ, the object of interest, we may neglect it for now and write

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

What does this expression tell us?

- $p(\theta)$ is the **prior density** which summarizes our knowledge about θ before knowing the data.
- $p(\theta|y)$ is the **posterior density** which summarizes our knowledge about θ once knowing the data.
- $p(y|\theta)$ is the **likelihood function** (well-known to us) that describes the distribution of the data given θ.

Objective of Bayesian econometrics

Use data to update our knowledge of an unknown parameter.

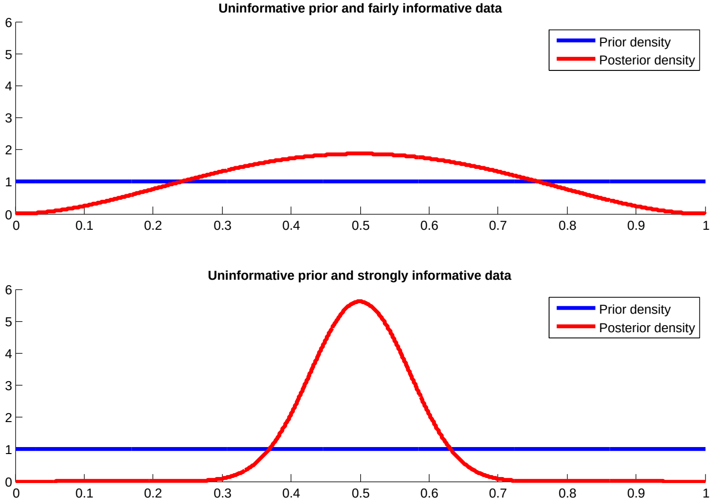
If the data are informative about the parameter, the posterior density should be less dispersed than the prior density: we "sharpen" our knowledge.

Example:

- We are interested in the Calvo price-adjustment parameter of a new Keynesian DSGE model. Let us assume all we know a priorily is that it is somewhere between 0 and 1. Then our prior density is simply

$$p(\theta) = 1, \qquad 0 \leq \theta \leq 1.$$

- Now let us add data. While the computation step in between might be quite involved and time-consuming, the posterior density might look like the ones on the next slide.



$$\text{Upper panel: } 90 \text{ percent interval} = [0.19, 0.81]$$

$$\text{Lower panel: } 90 \text{ percent interval} = [0.38, 0.62]$$

Use of the posterior density

We can use the posterior density to find

- a particularly “likely” parameter value (mean, median or mode), or
- an interval of parameter values that receive a high probability (e.g., 90 percent).

In the example above (lower panel), we found

- mean and mode $= 0.5$ (this is the value you would tell Ms. Lagarde or use if you simulated a monetary policy measure), and
- 90 percent interval $= [0.38, 0.62]$ (this is the range of values you would consider as a robustness check when simulating the monetary policy measure).

Use of the posterior density

Highest posterior density interval

Suppose you have found the posterior density $p(\theta|y)$. A $100(1-\alpha)\%$ credible interval is defined as

$$p(a \leq \theta \leq b|y) = \int_a^b p(\theta|y)d\theta = 1-\alpha.$$

The $100(1-\alpha)\%$ highest posterior density interval is the shortest $100(1-\alpha)\%$ credible interval.

Example: If the posterior is standard normal, then $[-1.96, 1.96]$, $[-1.75, 2.33]$, and $[-1.64, \infty]$ are all 95% credible intervals but only $[-1.96, 1.96]$ is the 95% highest posterior density interval.

Example: Bernoulli distribution

Taken from problem set 2

Let $y = (y_1, \ldots, y_N)'$ be a random sample from the Bernoulli distribution

$$p(y_i|\theta) = \begin{cases} \theta^{y_i}(1-\theta)^{1-y_i} & \text{if } y_i = 0 \text{ or } 1 \\ 0 & \text{otherwise} \end{cases}$$

This means we have binary data. We are interested in estimating the "success" probability θ. In classical statistics we would use the sample mean as a consistent estimator.

Bayesian estimation requires a prior distribution that summarizes our "pre-data" knowledge. Let us assume we have none. Thus we use the uniform prior

$$p(\theta) = 1, \qquad 0 \leq \theta \leq 1.$$

Questions: What is the posterior density? What is $\mathrm{E}(\theta|y)$?

To obtain

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

we first have to find (a) the likelihood $p(y|\theta)$ and (b) the marginal density of the data $p(y)$.

(a) Likelihood: Since we have a random sample, the y_i's are independent so that

$$p(y|\theta) = \prod_{i=1}^{N} p(y_i|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_i y_i}(1-\theta)^{N-\sum_i y_i}$$

(b) Marginal density of the data:

$$p(y) = \int_0^1 p(y,\theta)d\theta = \int_0^1 p(y|\theta)p(\theta)d\theta = \int_0^1 \theta^{\sum_i y_i} (1-\theta)^{N-\sum_i y_i} \times 1 \; d\theta$$

The integrand looks like the kernel of a Beta density in θ. Recall: the Beta(a, b) density uses the beta function $B(a, b)$ and is defined as

$$f_{Beta}(y) = B(a,b)^{-1} y^{a-1}(1-y)^{b-1}, \qquad 0 \leq y \leq 1.$$

Let us re-write the marginal density accordingly:

$$p(y) = \int_0^1 \theta^{1+\sum_i y_i - 1} (1-\theta)^{1+N-\sum_i y_i - 1} \, d\theta$$

This is the kernel of a Beta$(1 + \sum_i y_i, 1 + N - \sum_i y_i)$ density. Add the integrating constant:

$$p(y) = \underbrace{B(1+\textstyle\sum_i y_i, 1+N-\sum_i y_i)}_{\text{Remainder term } A \text{ independent of } \theta} \underbrace{\int_0^1 \frac{\theta^{1+\sum_i y_i - 1} (1-\theta)^{1+N-\sum_i y_i - 1}}{B(1+\sum_i y_i, 1+N-\sum_i y_i)} \, d\theta}_{\text{Beta pdf, integrates to 1}}$$

(c) Find the posterior:

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

Note that above we found that $p(y|\theta)p(\theta)$ can be written as

$$p(y|\theta)p(\theta) = \underbrace{A}_{\text{Remainder term independent of } \theta} \times \underbrace{pdf(\theta)}_{\text{pdf that integrates to 1.}}$$

Hence the posterior is

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} = \frac{A \times pdf(\theta)}{A \times \int pdf(\theta)d\theta} = \frac{A \times pdf(\theta)}{A} = pdf(\theta).$$

This is a general result. It implies that finding $p(y)$ yields the constant A that also shows up in the nominator. Hence, in practice we do not have calculate $p(y) = A$. All we need is to separate the nominator, $p(y|\theta)p(\theta)$, into a constant that is independent of θ and a pdf in θ. This pdf is the posterior $p(\theta|y)$ we are looking for.

Back to our example. Knowing $p(y)$ we substitute and rearrange:

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{\theta^{\sum_i y_i}(1-\theta)^{N-\sum_i y_i}}{B(1+\sum_i y_i, 1+N-\sum_i y_i)} \\ &= \frac{\theta^{1+\sum_i y_i-1}(1-\theta)^{1+N-\sum_i y_i-1}}{B(1+\sum_i y_i, 1+N-\sum_i y_i)}. \end{aligned}$$

Hence, the posterior is a Beta$(1+\sum_i y_i, 1+N-\sum_i y_i)$ density.

Alternatively, without knowing $p(y)$ we start from

$$p(\theta|y) \propto p(y|\theta)p(\theta) = \theta^{\sum_i y_i} (1-\theta)^{N-\sum_i y_i}$$

and conclude that this is the kernel of a Beta$(1+\sum_i y_i, 1+N-\sum_i y_i)$ density. Rearranging yields

$$p(\theta|y) \propto p(y|\theta)p(\theta) = B(1+\sum_i y_i, 1+N-\sum_i y_i) \times \frac{\theta^{1+\sum_i y_i - 1} (1-\theta)^{1+N-\sum_i y_i - 1}}{B(1+\sum_i y_i, 1+N-\sum_i y_i)}.$$

Since the second term is a pdf in θ (that integrates to 1), the first term must be $p(y)$ which cancels out when dividing $p(y|\theta)p(\theta)$ by $p(y)$. Therefore,

$$p(\theta|y) \sim Beta(1+\sum_i y_i, 1+N-\sum_i y_i).$$

Not surprisingly, we obtain the same result.

Point estimation: A possible point estimator is $\mathrm{E}(\theta|y)$. Since the mean of a Beta(a, b) distribution is $\frac{a}{a+b}$, we obtain

$$\mathrm{E}(\theta|y) = \frac{1 + \sum_i y_i}{1 + \sum_i y_i + 1 + N - \sum_i y_i} = \frac{1 + \sum_i y_i}{2 + N}$$

What does it mean?

- Without sample ($N = 0$, $\sum_i y_i = 0$), we have $\mathrm{E}(\theta|y) = 0.5$ which is the mean of the prior distribution.
- As the sample size increases, the sample information dominates more and more. To see this write

  $$\mathrm{E}(\theta|y) = \frac{1/N + \bar{y}}{2/N + 1}$$

  Hence the relevance of the prior information vanishes with increasing N. Eventually, as $N \to \infty$, $\mathrm{E}(\theta|y) \to \bar{y}$, the frequentist estimator.

Summary: How to find the posterior
(Assumption: random sampling)

- You are given the prior, $p(\theta)$, and the conditional distribution of a single observation, $p(y_i|\theta)$.
- Find the likelihood as
$$p(y|\theta) = \prod_{i=1}^{N} p(y_i|\theta).$$
- Write down the product of likelihood and prior,
$$p(y|\theta)p(\theta)$$
and separate it into a constant that is independent of θ and a pdf in θ.
- The constant is $p(y)$.
- The pdf in θ is the posterior $p(\theta|y)$

2. Natural Conjugate Priors

Definition

Conjugate prior: a prior distribution, when combined with the likelihood, that yields a posterior distribution which falls in the same class of distributions

Natural conjugate prior: a conjugate prior that has the same functional form as the likelihood

Natural conjugate priors are very helpful to find the posterior analytically. They also have a nice interpretation: since they have the same form as the likelihood, they can in fact be interpreted as a likelihood of a fictitious prior data sample.

Idea: Posterior is combination of

- fictitious prior data sample $\rightarrow$ prior distribution (= prior "likelihood"), and
- actual data sample $\rightarrow$ likelihood

Example: Bernoulli distribution with conjugate Beta prior

Let $y = (y_1, \ldots, y_N)'$ be a random sample from a Bernoulli distribution with unknown success probability θ and pmf

$$p(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}, \qquad y_i = 0 \text{ or } 1.$$

As seen above, this gives rise to the likelihood

$$p(y|\theta) = \prod_{i=1}^{N} p(y_i|\theta) = \theta^m (1-\theta)^{N-m},$$

where $m = \sum_{i=1}^{N} y_i$ is the number of successes.

Now suppose prior beliefs concerning θ are represented by a beta distribution:

$$p(\theta|\underline{\alpha}, \underline{\delta}) = B(\underline{\alpha}, \underline{\delta})^{-1} \theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\delta}-1}, \qquad 0 < \theta < 1.$$

The prior has the same functional form as the likelihood (the constant factor $B(\underline{\alpha}, \underline{\delta})^{-1}$ is irrelevant as it does not depend on θ and thus is part of the remainder term).

Interpretation of the prior: think of a "prior sample" of size P drawn from the Bernoulli distribution and suppose this sample contains $\underline{\alpha} - 1$ successes and $\underline{\delta} - 1$ non-successes.

Then $P = \underline{\alpha} + \underline{\delta} - 2$ and the "prior likelihood" is

$$p_{\text{prior}}(y|\theta) = \theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\delta}-1},$$

which is proportional to the beta distribution used as a prior above.

Hence, we may interpret the beta prior as knowledge gained from an imaginary prior sample of size $P = \underline{\alpha} + \underline{\delta} - 2$ with $\underline{\alpha} - 1$ successes.

Let us find the posterior to strengthen this interpretation. We have (up to constants) the prior

$$p(\theta|\underline{\alpha}, \underline{\delta}) \propto \theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\delta}-1}$$

and the likelihood

$$p(y|\theta) \propto \theta^{m}(1-\theta)^{N-m}.$$

Multiplying likelihood and prior yields

$$p(\theta|y) \propto \theta^{m}(1-\theta)^{N-m} \times \theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\delta}-1} = \underbrace{\theta^{m+\underline{\alpha}-1}(1-\theta)^{N-m+\underline{\delta}-1}}_{\text{kernel of a Beta}(m+\underline{\alpha},\, N-m+\underline{\delta})\text{ distribution}}$$

We conclude that the posterior is a Beta$(m+\underline{\alpha}, N-m+\underline{\delta})$ distribution:

$$p(\theta|y) = B(m+\underline{\alpha}, N-m+\underline{\delta})^{-1}\theta^{m+\underline{\alpha}-1}(1-\theta)^{N-m+\underline{\delta}-1}.$$

As a point estimator, let us use the expectation of the posterior which is

$$\mathrm{E}(\theta|y) = \frac{m+\underline{\alpha}}{m+\underline{\alpha}+N-m+\underline{\delta}} = \frac{m+\underline{\alpha}}{N+\underline{\alpha}+\underline{\delta}}.$$

To facilitate the "prior sample" interpretation, substitute $P+2=\underline{\alpha}+\underline{\delta}$ and recall that $\alpha-1$ is the number of "prior sample" successes:

$$\mathrm{E}(\theta|y) = \frac{m+\underline{\alpha}-1+1}{N+P+2} = \frac{\#\text{ sample successes} + \#\text{ prior sample successes} + 1}{\text{sample size} + \text{prior sample size} + 2}.$$

3. Assessing the Influence of the Prior

Dependency on the prior

We have seen that the choice of the prior partly determines the posterior result.

This is intended because the prior reflects our prior (subjective) knowledge.

But sometimes we just do not have any good prior information or we do not want to use it because others may not share our view. What can we do in such a case?

Prior sensitivity analysis

The first way to check by how much our prior drives the posterior results is to use a prior sensitivity analysis. This is, in principle, straightforward:

- Identify all sensible priors. Example 1: choose the P priors that reflect the P different views put forward in the literature. Example 2: choose a fine grid of P priors from a sensible range.
- Find posterior means, e.g., $\bar{\mu}^{(1)}, \ldots, \bar{\mu}^{(P)}$, based on the P priors.
- State the minimum and maximum of all posterior means. This is called extreme bounds analysis. If the data are very informative, then a result may survive all priors.

Example: Bernoulli distribution with conjugate Beta prior

Prior sensitivity analysis: consider a range of beta priors for the success probability.

Here:

- Range of modi $\underline{\theta}^{(i)} = \frac{\underline{\alpha}^{(i)}-1}{\underline{\alpha}^{(i)}+\underline{\delta}^{(i)}-2} = \frac{\text{\# prior sample successes}}{\text{sample size}}$ between 0.1 and 0.9
- Constant prior sample size $P = \underline{\alpha}^{(i)} + \underline{\delta}^{(i)} - 2 = 10$
- This yields parameters

$$\underline{\alpha}^{(i)} = \underline{\theta}^{(i)} P + 1 \qquad \Rightarrow \qquad \underline{\alpha}^{(i)} \in [2, \ldots, 10]$$

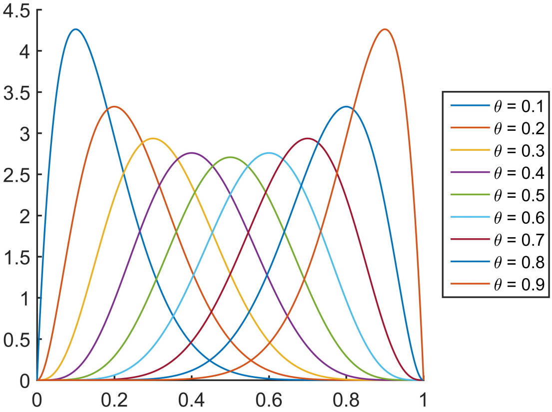
and

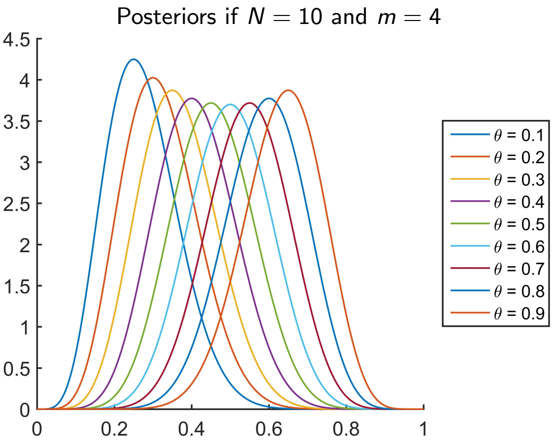
$$\underline{\delta}^{(i)} = (1 - \underline{\theta}^{(i)}) P + 1 \qquad \Rightarrow \qquad \underline{\delta}^{(i)} \in [10, \ldots, 2]$$

- The posterior means are

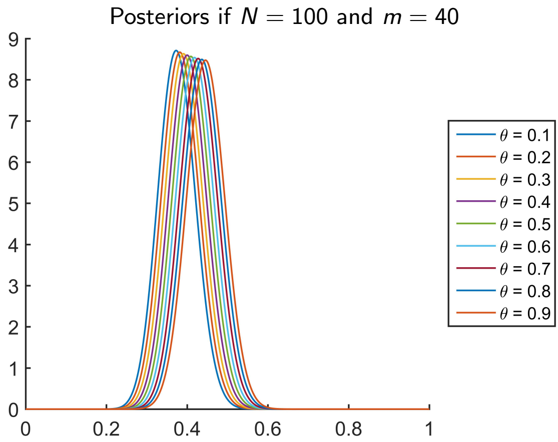
$$\mathrm{E}(\theta^{(i)}|y) = \frac{m + \underline{\alpha}^{(i)}}{N + P + 2}$$

Priors if $P = 10$ and modus $\underline{\theta}^{(i)} \in [0.1, \ldots, 0.9]$





Posterior means: $\mathrm{E}(\theta^{(i)}|y) \in [0.27, \dots, 0.64]$



Posterior means: $E(\theta^{(i)}|y) \in [0.38, \ldots, 0.45]$

Noninformative prior

Alternatively, choose a prior that is not informative relative to the sample information.

In the Bernoulli example, this means to choose $\underline{\alpha}$ and $\underline{\delta}$ near 1. Then the beta distribution differs only slightly from a uniform distribution.

A fully noninformative prior sets $\underline{\alpha} = \underline{\delta} = 1$ so that the beta distribution degenerates to a uniform distribution which places the same weight on every possible success probability. The posterior mean becomes

$$\mathrm{E}(\theta|y) = \frac{m + \underline{\alpha}}{N + \underline{\alpha} + \underline{\beta}} = \frac{m+1}{N+2}.$$

4. Model Comparison

Competing models

In econometrics we often have competing models. Which one should we trust most?

Suppose we have m models, $M_1, \dots, M_m$. Each model M_i depends on a specific set of parameters θ^i.

Then each model has a prior, posterior, and likelihood:

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

But how can we discriminate between the models?

Model probabilities

Let us apply Bayesian econometrics to model selection—update prior beliefs by using data:

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

What does this expression tell us?

- $p(M_i)$ is the **prior model probability** which summarizes our knowledge about M_i before knowing the data (and is discrete now).
- $p(M_i|y)$ is the **posterior model probability** which summarizes our knowledge about M_i once knowing the data.
- $p(y|M_i)$ is the **marginal likelihood** that describes the distribution of the data given M_i. It does not depend θ because we need a relationship between model and data alone. We show on the next slide how to obtain it.

Marginal likelihood

How can we find the marginal likelihood of model M_i?

The relationship between model and data, $p(y|M_i)$, we are looking for can be interpreted as an average relationship. For each model M_i there is a set of possible parameter values θ^j. For each θ^j we have the conventional likelihood

$$p(y|\theta^j, M_i).$$

To find the average relationship between model and data, recall statistics: integrate across all possible θ^j (here: continuous case):

$$p(y|M_i) = \int \underbrace{p(y|\theta^j, M_i)}_{\text{conditional p.}} \underbrace{p(\theta^j|M_i)}_{\text{marginal p.}} d\theta^j.$$

While it may be difficult to do the integration in practice, it is straightforward in principle.

We get the same result if we start from

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

bring $p(y|M_i)$ to the lhs and integrate over all θ^i;

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

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

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

because $\int p(\theta^i|y, M_i)d\theta^i = 1$, like every density.

Posterior odds ratio

Since it is often difficult to find the denominator in

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

we often compute the posterior odds ratio

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

If the set of models is exhaustive (such that we have one more equation)

$$p(M_i|y) + \ldots + p(M_m|y) = 1,$$

the posterior odds ratio can be used to compute the posterior model probabilities.

Example: $m = 2$.

Then knowing

$$PO_{12} = \frac{p(M_1|y)}{p(M_2|y)} \quad \text{and} \quad p(M_1|y) + p(M_2|y) = 1$$

is sufficient to compute

$$p(M_1|y) = PO_{12}\, p(M_2|y) = PO_{12}[1 - p(M_1|y)] \quad \Rightarrow \quad p(M_1|y) = \frac{PO_{12}}{1 + PO_{12}}$$

and

$$p(M_2|y) = 1 - p(M_1|y) = \frac{1}{1 + PO_{12}}.$$

Bayes factor

Suppose prior model probabilities are equal: $p(M_1) = \ldots = p(M_m)$ such that the **prior odds ratio** is

$$\frac{p(M_i)}{p(M_j)} = 1.$$

This simplifies the posterior odds ratio to

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

which is simply the ratio of marginal likelihoods, also called the Bayes factor

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

5. Prediction

Suppose we want to predict some (future) unknown data y^* based on observed data y.

Bayesian econometrics: use the **predictive density** $p(y^*|y)$.

Unfortunately, the model for the random variable y^* depends on the parameters θ which are also unknown and treated as a random variable. Thus, we really have the joint density $p(y^*, \theta|y)$.

The predictive density is then obtained by integrating over all θ:

$$p(y^*|y) = \int p(y^*, \theta|y) d\theta.$$

The expression

$$p(y^*|y) = \int p(y^*, \theta|y) d\theta.$$

is inconvenient because it entails writing down a joint probability model for unknown data *and* unknown parameters (conditional on y).

It is simpler to think of a model of y^* conditional on knowing both the parameters and the data. This is achieved by a conditional-marginal factorization of the joint probability:

$$p(y^*|y) = \int p(y^*|\theta, y) p(\theta|y) d\theta.$$

It turns out that the integrand now consists of

- a model of y^* conditional on both the parameters and the data, and
- the posterior probability of θ conditional on the data (which is the outcome of Bayesian estimation).

6. Computation

The Bayesian "bottleneck"

The previous slides and equations is, conceptually, all you need to know about Bayesian econometrics.

Nice and simple, but how to do the computation unless we specialize in very simple models?

In particular, how to evaluate those integrals?

Unfortunately, we typically need numerical methods because analytical solutions are not available.

Where we need to integrate

Suppose θ is a vector of k elements, $\theta = (\theta_1, \ldots, \theta_k)'$. We need to solve a multivariate integral in the following cases (nonexhaustive examples):

- Point estimation: mean of the posterior density

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

- Measure of estimation uncertainty: variance (or standard deviation) of the posterior density

$$\mathrm{Var}(\theta_i|y) = \mathrm{E}(\theta_i^2|y) - [\mathrm{E}(\theta_i|y)]^2 = \int \theta_i^2 p(\theta|y) d\theta - \left[\int \theta_i p(\theta|y) d\theta\right]^2$$

- Probabilities: is the parameter θ_i in a certain interval, e.g., is it positive? Evaluate its marginal density

$$p(\theta_i > 0|y) = \int_0^\infty p(\theta_i|y) d\theta_i$$

The general integration problem

By suitable definition of a weight function $g(\theta)$, each of the previous problems and many more can be written as

$$E(g(\theta)|y) = \int g(\theta)p(\theta|y)d\theta.$$

For example,

- $g(\theta) = \theta_i$ yields $E(\theta_i|y)$
- $g(\theta) = \theta_i^2$ yields $E(\theta_i^2|y)$ (needed for the variance)
- $g(\theta) = 1(\theta_i > 0)$ yields $p(\theta_i > 0|y)$.

Note: $1(A)$ is an indicator function which equals 1 if A is true and 0 otherwise.

Monte Carlo integration

If no analytical solution is available, we can simulate the statistic of interest:

- Use the random number generator of a computer to obtain $s = 1, \ldots, S$ random draws $\theta^{(s)}$ from $p(\theta|y)$. This is called posterior simulation.
- For each draw calculate $g(\theta^{(s)})$.
- By a law of large numbers,

$$\hat{g}_S = \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \xrightarrow{\text{p}} \mathrm{E}(g(\theta)|y).$$

Hence, if we choose S large enough we can be confident that $\hat{g}_S$ is near the statistic of interest, $\mathrm{E}(g(\theta)|y)$.

But exactly how confident shall we be?

Approximation error

Under general conditions, a central limit theorem applies which states

$$\sqrt{S}\,[\hat{g}_S - \mathrm{E}(g(\theta)|y)] \xrightarrow{\mathrm{d}} N(0, \sigma_g^2),$$

where $\sigma_g^2 = \mathrm{Var}(g(\theta)|y)$.

Hence, for large number of replications S, it approximately holds that

$$\hat{g}_S - \mathrm{E}(g(\theta)|y) \sim N(0, \sigma_g^2/S),$$

from which an approximate $1-\alpha$ interval can be constructed (here $\alpha = 5$ percent):

$$\Pr\left[-1.96\sigma_g/\sqrt{S} \leq \hat{g}_S - \mathrm{E}(g(\theta)|y) \leq 1.96\sigma_g/\sqrt{S}\right] = 0.95.$$

Instead of reporting such an interval, you may just report the **numerical standard error**

$$\sigma_g/\sqrt{S}.$$

How many replications?

The numerical standard error is the ratio of σ_g to $\sqrt{S}$.

Recall: σ_g is the standard error of the posterior function $g(\theta)$.

A reliable simulation should thus have a standard error much smaller than σ_g.

Hence, we should choose S such that $\sigma_g/\sqrt{S}$ has an appropriate size. For example, to achieve a numerical standard error of as small as 1 percent of σ_g, we need $S = 10{,}000$ replications.

Implementation in Matlab

Matlab has a powerful but costly Statistics Toolbox. If you have access to it, perfect (currently, Kiel University has a campus licence).

If not, no problem: use the great toolbox of James P. LeSage. Download it for free here.

Gary Koop also offers random numbers, distributions etc needed for this course on the companion website. Parts are taken from LeSage's toolbox.

Execution time in Matlab

Matlab is fast if you work in matrices instead of loops.

Example: you want to get 1,000,000 draws from the standard normal distribution.

Good: `theta=randn(1000000,1);`

Bad: `for i = 1:1000000, theta(i)=randn(1,1); end`

The difference in computing time is huge: 0.021 versus 1.045 seconds (on my notebook). This is roughly the factor 50!

How to speed up your Matlab programs

- Use matrices instead of loops.
- If you nevertheless need to use a loop, make sure that computationally intensive tasks such as matrix inversion are executed as infrequently as possible.
- Follow the recommendations Matlab gives you in the Editor.
- Use the Matlab profiler—just type `help profile` to get help on it.
- If possible, use Matlab's parallel computing toolbox if tasks can be parallelized.