Chapter 7: Bayesian Econometrics

Christophe Hurlin

University of Orléans

June 26, 2014

Section 1

Introduction

1. Introduction

The outline of this chapter is the following:

Section 2. Prior and posterior distribution

Section 3. Posterior distributions and inference

Section 4. Applications (VAR and DSGE)

Section 4. Numerical simulations

1. Introduction

References

- Geweke J. (2005), Contemporary Bayesian Econometrics and Statistics. New York: John Wiley and Sons. (advanced)
- Geweke J., Koop G. and Van Dijk H. (2011), The Oxford Handbook of Bayesian Econometrics. Oxford University Press.
- Greene W. (2007), Econometric Analysis, sixth edition, Pearson - Prentice Hil
- Greenberg E. (2008), Introduction to Bayesian Econometrics, Cambridge University Press. **(recommended)**
- Koop, G. (2003), Bayesian Econometrics. New York: JohnWiley and Sons.
- Lancaster T. (2004), An Introduction to Modern Bayesian Inference. Oxford University Press.

1. Introduction

Notations: In this course, I will (try to...) follow some conventions of notation.

Y	random variable
y	realisation
$f_Y(y)$	probability density function or probability mass function
$F_Y(y)$	cumulative distribution function
$\Pr(.)$	probability
$\mathbf{y}$	vector
$\mathbf{Y}$	matrix

Problem: this system of notations does not allow to discriminate between a vector (matrix) of random elements and a vector (matrix) of non-stochastic elements (realisation).

Abadir and Magnus (2002), Notation in econometrics: a proposal for a standard, Econometrics Journal.

Section 2

Prior and posterior distributions

2. Prior and posterior distributions

Objectives

The objective of this section are the following:

1. Introduce the concept of **prior distribution**
2. Define the **hyperparameters**
3. Define the concept of **posterior distribution**
4. Define the concept of **unormalised posterior distribution**

2. Prior and posterior distributions

In statistics, there a distinction between two concepts of **probability**:

1. **Frequentist** probability
2. **Subjective** probability

2. Prior and posterior distributions

Frequentist probability

Frequentists restrict the assignment of probabilities to statements that describe the outcome of an experiment that can be **repeated**.

Example

Statement A1: "A coin tossed three times will come up heads either two or three times." We can imagine repeating the experiment of tossing a coin three times and recording the number of times that two or three heads were reported. Then:

$$\Pr(A_1) = \lim_{n \to \infty} \frac{\text{number of times two or three heads occurs}}{n}$$

2. Prior and posterior distributions

Subjective probability

1. Those who take the subjective view of probability believe that probability theory is applicable to any situation in which there is **uncertainty**.
2. Outcomes of repeated experiments fall in that category, but so do statements about tomorrow's weather, which are not the outcomes of repeated experiments.
3. Calling probabilities "subjective" does not imply that they can be set arbitrarily, and probabilities set in accordance with the axioms are consistent.

2. Prior and posterior distributions

Reminder

The probability of event A is denoted by $\Pr(A)$. Probabilities are assumed to satisfy the following axioms:

1. $0 \leq \Pr(A) \leq 1$
2. $\Pr(A) = 1$ if A represents a logical truth
3. If A and B describe disjoint events (events that cannot both occur), then $\Pr(A \cup B) = \Pr(A) + \Pr(B)$
4. Let $\Pr(A|\,B)$ denote "the probability of A, given (or conditioned on the assumption) that B is true." Then

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

2. Prior and posterior distributions

Reminder

Definition

The union of A and B is the event that A or B (or both) occur; it is denoted by $A \cup B$.

Definition

The intersection of A and B is the event that both A and B occur; it is denoted by $A \cap B$.

2. Prior and posterior distributions

Questions

1. What is the fundamental idea of the **posterior distribution**?
2. How it can be computed from the **likelihood function** and the **prior distribution**?

2. Prior and posterior distributions

Example (Subjective view of probability)

Let Y a binary variable with $Y = 1$ if a coin toss results in a head and 0 otherwise, and let

$$\Pr(Y = 1) = \theta$$

$$\Pr(Y = 0) = 1 - \theta$$

which is assumed to be constant for each trial. In this model, θ is a **parameter** and the value of Y is the **data (realisation y)**.

2. Prior and posterior distributions

Example (cont'd)

Under these assumptions, Y is said to have the Bernoulli distribution, written as

$$Y \sim Be(\theta)$$

with a probability mass function (pmf)

$$f_Y(y;\theta) = \Pr(Y = y) = \theta^y (1-\theta)^{1-y}$$

We consider a sample of $i.i.d.$ variables $(Y_1, .., Y_n)$ that corresponds to n repeated experiments. The realisation is denoted by $(y_1, .., y_n)$.

2. Prior and posterior distributions

Frequentist approach

1. From the **frequentist point of view**, probability theory can tell us something about the **distribution of the data for a given** θ.

$$Y \sim Be(\theta)$$

2. The parameter θ is an unknown number between zero and one.

3. It is not given a **probability distribution of its own**, because it is not regarded as being the **outcome of a repeated experiment**.

2. Prior and posterior distributions

Fact (Frequentist approach)

In a frequentist approach, the parameters θ are considered as **constant terms** *and the aim is to study the distribution of the data given θ, through the likelihood of the sample.*

Reminder

The likelihood function is defined as to be:

$$L_n : \Theta \times \mathbb{R}^n \rightarrow \mathbb{R}^+$$

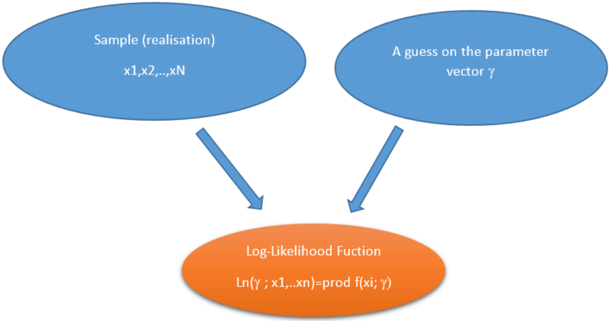
$$(\theta; y_1, .., y_n) \longmapsto L_n (\theta; y_1, .., y_n) = f_{Y_1,..,Y_N} (y_1, y_2, .., y_n; \theta)$$

Under the $i.i.d.$ assumption

$$(\theta; y_1, .., y_n) \longmapsto L_n (\theta; y_1, .., y_n) = \prod_{i=1}^{n} f_Y (y_i; \theta)$$

2. Prior and posterior distributions

Remark: the (log-)likelihood function depends on two arguments: the sample (realisation) and θ



2. Prior and posterior distributions

Reminder

In our example, Y is a discrete variable and the likelihood can be interpreted as the joint probability to observe the sample (realisation) $(y_1, .., y_n)$ given a value of θ

$$L_n(\theta; y_1.., y_n) = \Pr((Y_1 = y_1) \cap ... \cap (Y_n = y_n))$$

If the sample $(Y_1, .., Y_n)$ is $i.i.d.$, then:

$$L_n(\theta; y_1.., y_n) = \prod_{i=1}^{n} \Pr(Y_i = y_i) = \prod_{i=1}^{n} f_Y(y_i; \theta)$$

Reminder

In our example, the likelihood of the sample $(y_1, .., y_n)$ is

$$\begin{aligned} L_n(\theta; y_1 .., y_n) &= \prod_{i=1}^{n} f_Y(y_i; \theta) \\ &= \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \end{aligned}$$

Or equivalently

$$L_n(\theta; y_1 .., y_n) = \theta^{\Sigma y_i} (1-\theta)^{\Sigma(1-y_i)}$$

2. Prior and posterior distributions

Notations:

In the rest of the chapter, I will use the following alternative notations:

$$L_n(\theta; y) \equiv L(\theta; y_1, .., y_n) \equiv L_n(\theta)$$

$$\ell_n(\theta; y) \equiv \ln L_n(\theta; y) \equiv \ln L(\theta; y_1, .., y_n) \equiv \ln L_n(\theta)$$

Frequentist approach

- From the subjective point of view, however, θ is an **unknown quantity**.
- Since there is uncertainty over its value, it can be regarded as a **random variable** and assigned a probability distribution.
- Before seeing the data, it is assigned a **prior distribution**

$$\pi(\theta) \text{ with } 0 \leq \theta \leq 1$$

2. Prior and posterior distributions

Definition (Prior distribution)

In a Bayesian framework, the parameters θ associated to the distribution of the data, are considered as **random variables**. Their distribution is called the **prior distribution** and is denoted by $\pi(\theta)$.

2. Prior and posterior distributions

Remark

In our example, the endogenous variable Y_i is discrete (0 or 1):

$$Y_i \sim Be(\theta)$$

but the parameter $\theta = \Pr(Y_i = 1)$ can be considered as a continuous random variable over $[0, 1]$: in this case $\pi(\theta)$ is a pdf.

2. Prior and posterior distributions

Example (Prior distribution)

For instance, we may postulate that:

$$\theta \sim U_{[0,1]}$$

2. Prior and posterior distributions

Remark

Whatever the type of the distribution of the endogenous variable (discrete or continuous), the prior distribution is generally **continuous**.

$$\pi(\theta) = \text{probability density function (pdf)}$$

2. Prior and posterior distributions

Definition (Uninformative prior)

An **uninformative** prior is a flat prior. Example: $\theta \sim U_{[0,1]}$

2. Prior and posterior distributions

Remark

In most of cases, the prior distribution is parametrised, *i.e.* the pdf $\pi(\theta; \gamma)$ depends on a set of parameters γ.

Definition (Hyperparameters)

The parameters of the prior distribution, called **hyperparameters**, are supplied by the researcher.

2. Prior and posterior distributions

Example (Hyperparameters)

If $\theta \in \mathbb{R}$ and if the prior distribution is normal

$$\pi(\theta; \gamma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)$$

with $\gamma = \left(\mu, \sigma^2\right)^\top$ the vector of hyperparameters.

2. Prior and posterior distributions

Example (Beta prior distribution)

If $\theta \in [0,1]$, a common (parametrised) prior distribution is the **Beta distribution** denoted $B(\alpha, \beta)$.

$$\pi(\theta; \gamma) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \qquad \alpha, \beta > 0 \quad \theta \in [0,1]$$

with $\gamma = (\alpha, \beta)^{\top}$ the vector of hyperparameters.

Beta distribution: reminder

The **gamma function**, denoted $\Gamma(p)$, is defined as to be:

$$\Gamma(p) = \int_0^{+\infty} e^{-x} x^{p-1} dx \qquad p > 0$$

with:

$$\Gamma(p) = (p-1)\,\Gamma(p-1)$$

$$\Gamma(p) = (p-1)! = (p-1) \times (p-2) \ldots \times 2 \times 1 \quad \text{if } p \in \mathbb{N}$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

.

2. Prior and posterior distributions

The **Beta distribution** $B(\alpha, \beta)$ has very interesting features:

1. Depending on the choice of α and β, this prior can capture beliefs that indicate θ is centered at $1/2$, or it can shade θ toward zero or one;
2. It can be highly concentrated, or it can be spread out;
3. When both parameters are less than one, it can have two modes.

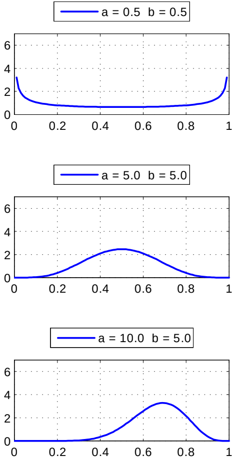
The shape of a **Beta distribution** can be understood by examining its mean and variance:

$$\theta \sim B(\alpha, \beta)$$

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha + \beta} \qquad \mathbb{V}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

1. The mean is equal to $1/2$ if $\alpha = \beta$
2. A larger α (β) shades the mean toward 1 (0)
3. the variance decreases as α or β increases.

2. Prior and posterior distributions



2. Prior and posterior distributions

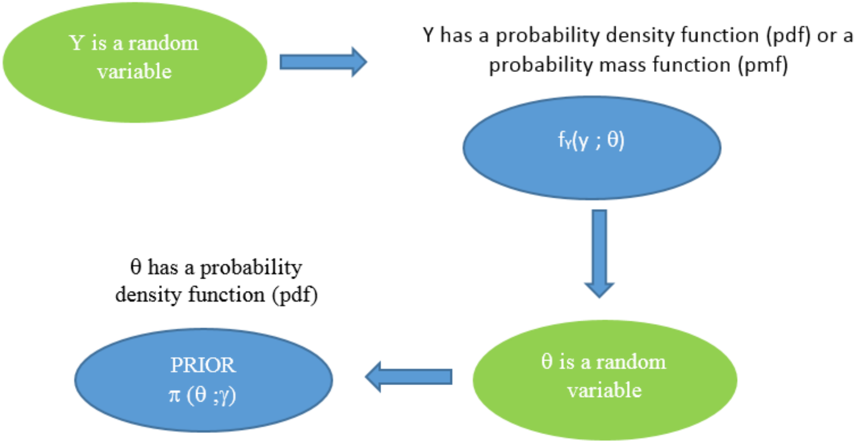
```
%=====================================================
% PURPOSE: Reproduce the Figure 2 of the Chapter 7
% Lecture: "Advanced Econometrics"
%-------------------------
% Author: Christophe Hurlin, University of Orleans
% Version: v1. June 2014
%=====================================================

clear all ; clc ; close all

theta=(0:0.01:1)';      % Create a vector of values for theta (0,1)
a=[0.5 1 5 30 10 1]';   % Values for alpha
b=[0.5 1 5 30 5 30]';   % Values for beta

figure
for i=1:6
    subplot (3,2,i)
    plot(theta,betapdf(theta,a(i),b(i)),'LineWidth',1.5)
    legend(sprintf('a = %1.1f  b = %1.1f',a(i),b(i)),'Location','NorthOutside' )
    grid ('on')
    ylim([0 7])
end
```

2. Prior and posterior distributions



2. Prior and posterior distributions

Remark

- In some models, the **hyperparameters** are **stochastic**: as for the parameters of interest θ, they have a distribution.
- This models are called **hierarchical models**

2. Prior and posterior distributions

Definition (Hierarchical models)

In a **hierarchical model**, we add one or more additional levels, where the hyperparameters themselves are given a prior distribution depending on another set of hyperparameters.

Example (hierarchical model)

An example of hierarchical model is given by

$$y : \text{pdf } f\left(y|\,\theta\right)$$

$$\theta : \text{pdf } \pi\left(\theta|\,\alpha\right)$$

$$\alpha : \text{pdf } \pi\left(\alpha|\,\beta\right)$$

where the hyperparameters β are constant terms.

2. Prior and posterior distributions

Definition (Posterior distribution)

Bayesian inference centers on the **posterior distribution** $\pi(\theta|\, y)$, which is the **conditional distribution** of the random variable θ given the data (realisation of the sample) $y = (y_1, .., y_n)$.

$$\theta|\, (Y_1 = y_1, .. Y_n = y_n) \sim \text{posterior distribution}$$

2. Prior and posterior distributions

Remark

When there is more than one parameter, the posterior distribution is a joint conditional distribution of all the parameters given the observed data.

2. Prior and posterior distributions

Notations

$\pi(\theta)$	prior distribution $=$ pdf of θ
$\pi(\theta \mid y)$	posterior distribution $=$ conditional pdf
	Discrete endogenous variable
$p(y;\theta)$	probability mass function (pmf)
$\Pr(A)$	probability of event A
	Continuous endogenous variable
$f_Y(y;\theta)$	probability distribution function (pdf)
$F_Y(y;\theta)$	cumulative distribution function

2. Prior and posterior distributions

Warning

For all the general definitions and the general results, we will employ the notation $f_Y(y;\theta)$ for both probability mass and density functions

Be careful with the interpretation of $f_Y(y;\theta)$: density (continuous case) or directly probability (discrete case).

2. Prior and posterior distributions

Example (Discrete random variable)

If $Y \sim Be(\theta)$, the pmf given by:

$$f_Y(y;\theta) = \Pr(Y = y) = \theta^y (1-\theta)^{1-y}$$

is a probability.

2. Prior and posterior distributions

Example (Continuous random variable)

If $Y \sim N\left(\mu, \sigma^2\right)$ with $\theta = \left(\mu, \sigma^2\right)^\top$, the pdf

$$f_Y\left(y; \theta\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y-\mu\right)^2}{2\sigma^2}\right)$$

is not a probability. The probability is given by

$$\Pr\left(Y \leq y\right) = F_Y\left(y; \theta\right) = \int_{-\infty}^{y} f_Y\left(x; \theta\right) dx$$

$$\Pr\left(Y = y\right) = 0$$

2. Prior and posterior distributions

The posterior density function $\pi(\theta | y)$ is computed by **Bayes theorem**:

Theorem (Bayes's Theorem)

For events A and B, the conditional probability of event A given that B has occurred is

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

2. Prior and posterior distributions

Bayes theorem:

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

By setting $A = \theta$ and $B = y$, we have:

$$\pi(\theta|y) = \frac{f_{Y|\theta}(y|\theta) \times \pi(\theta)}{f_Y(y)}$$

where

$$f_Y(y) = \int f_{Y|\theta}(y|\theta)\,\pi(\theta)\,d\theta$$

2. Prior and posterior distributions

Definition (Posterior distribution)

For one observation y_i, the **posterior distribution** is the conditional distribution of θ given y_i, defined as to be:

$$\pi\left(\theta \mid y_i\right) = \frac{f_{Y_i|\theta}\left(y_i \mid \theta\right) \times \pi\left(\theta\right)}{f_{Y_i}\left(y_i\right)}$$

where

$$f_{Y_i}\left(y_i\right) = \int_{\Theta} f_{Y_i|\theta}\left(y_i \mid \theta\right) \pi\left(\theta\right) d\theta$$

and Θ the support of the distribution of θ.

2. Prior and posterior distributions

Remark

$$\pi\left(\theta \mid y_i\right) = \frac{f_{Y_i|\theta}\left(y_i \mid \theta\right) \times \pi\left(\theta\right)}{f_{Y_i}\left(y_i\right)}$$

The term $f_{Y_i|\theta}\left(y_i \mid \theta\right)$ corresponds to the likelihood of the observation y_i :

$$f_{Y_i|\theta}\left(y_i \mid \theta\right) = L_i\left(\theta; y_i\right)$$

2. Prior and posterior distributions

Remark

The effect of dividing by $f_{Y_i}(y_i)$ is to make $\pi(\theta | y_i)$ a normalized probability distribution: integrating $\pi(\theta | y_i)$ with respect to θ yields 1:

$$\begin{aligned} \int \pi(\theta | y_i)\, d\theta &= \int \frac{f_{Y_i|\theta}(y_i | \theta) \times \pi(\theta)}{f_{Y_i}(y_i)} d\theta \\ &= \frac{1}{f_{Y_i}(y_i)} \int f_{Y_i|\theta}(y_i | \theta) \times \pi(\theta)\, d\theta \\ &= 1 \end{aligned}$$

since $f_{Y_i}(y_i) = \int_{\Theta} f_{Y_i|\theta}(y_i | \theta)\, \pi(\theta)\, d\theta$.

Discrete variable

For a discrete random variable Y_i, by setting $A = \theta$ and $B = y$, we have:

$$\underbrace{\pi\left(\theta | y_i\right)}_{\text{posterior (pdf)}} = \frac{\underbrace{p\left(y_i | \theta\right)}_{\text{cond. probability}} \times \underbrace{\pi\left(\theta\right)}_{\text{prior (pdf)}}}{\underbrace{p\left(y_i\right)}_{\text{probability}}}$$

where

$$p\left(y_i\right) = \int p\left(y_i | \theta\right) \pi\left(\theta\right) d\theta$$

2. Prior and posterior distributions

Definition (Prior predictive distribution)

The term $p(y_i)$ or $f_{Y_i}(y_i)$ is sometimes called the **prior predictive distribution**

$$p(y_i) = \int p(y_i|\theta)\,\pi(\theta)\,d\theta$$

$$f_{Y_i}(y_i) = \int f_{Y|\theta}(y_i|\theta)\,\pi(\theta)\,d\theta$$

2. Prior and posterior distributions

Remark

1. In general, we consider an $i.i.d.$ sample $Y = (Y_1, .., Y_N)$ with a realisation (data) $y = (y_1, .., y_n)$, and not only one observation.

2. In this case, the posterior distribution can be written as function of the likelihood of the $i.i.d.$ sample $y = (y_1, .., y_n)$.

$$L_n(\theta; y_1, .., y_n) = \prod_{i=1}^{n} L_i(\theta; y_i) = \prod_{i=1}^{n} f_{Y_i}(y_i)$$

2. Prior and posterior distributions

Definition (Posterior distribution, sample)

For sample $(y_1, .., y_n)$, the **posterior distribution** is the conditional distribution of θ given y_i, defined as to be:

$$\pi(\theta | y_1, ..y_n) = \frac{L_n(\theta; y_1, .., y_n) \times \pi(\theta)}{f_{Y_1,..,Y_n}(y_1, .., y_n)}$$

where $L_n(\theta; y_1, .., y_n)$ is the likelihood of the sample and

$$f_{Y_1,..,Y_n}(y_1, .., y_n) = \int_{\Theta} L_n(\theta; y_1, .., y_n)\, \pi(\theta)\, d\theta$$

and Θ the support of the distribution of θ.

2. Prior and posterior distributions

Notations

$$\pi\left(\theta \mid y_1, ..y_n\right) = \frac{L_n\left(\theta; y_1, .., y_n\right) \times \pi\left(\theta\right)}{f_{Y_1,..,Y_n}\left(y_1, .., y_n\right)}$$

For simplicity, we skip the notation $\left(y_1, .., y_n\right)$ for the sample and we put only the generic term y :

$$\pi\left(\theta \mid y\right) = \frac{L_n\left(\theta; y\right) \times \pi\left(\theta\right)}{f_Y\left(y\right)} = \frac{f_{Y|\theta}\left(y \mid \theta\right) \times \pi\left(\theta\right)}{f_Y\left(y\right)}$$

2. Prior and posterior distributions

Remark

$$\pi\left(\theta \middle| y_1, ..y_n\right) = \frac{L_n\left(\theta; y_1, .., y_n\right) \times \pi\left(\theta\right)}{f_{Y_1,..Y_n}\left(y_1, .., y_n\right)}$$

In this setting, the data $\left(y_1, ..y_n\right)$ are viewed as constants whose (marginal) distributions do not involve the parameters of interest θ.

$$f_{Y_1,..Y_n}\left(y_1, .., y_n\right) = \text{constant}$$

2. Prior and posterior distributions

Remark (cont'd)

As a consequence, the Bayes theorem

$$\Pr\left(\text{parameters} \middle| \text{data}\right) = \frac{\Pr\left(\text{data} \middle| \text{parameters}\right) \times \Pr\left(\text{parameters}\right)}{\Pr\left(\text{data}\right)}$$

implies that

$$\underbrace{\Pr\left(\text{parameters} \middle| \text{data}\right)}_{\text{Posterior Density}} \propto \underbrace{\Pr\left(\text{data} \middle| \text{parameters}\right)}_{\text{Likelihood}} \times \underbrace{\Pr\left(\text{parameters}\right)}_{\text{Prior Density}}$$

where the symbol "$\propto$" means "is proportional to."

2. Prior and posterior distributions

Definition (Unormalised posterior distribution)

The **unormalised posterior distribution** is the product of the likelihood of the sample and the prior distribution:

$$\pi\left(\theta \middle| y_1, ..y_n\right) \propto L_n\left(\theta; y_1, .., y_n\right) \times \pi\left(\theta\right)$$

or with simplest notations

$$\pi\left(\theta \middle| y\right) \propto L_n\left(\theta; y\right) \times \pi\left(\theta\right)$$

where the symbol "$\propto$" means "is proportional to."

2. Prior and posterior distributions

Example (Posterior distribution)

Consider an $i.i.d.$ sample $(Y_1, .., Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$ and:

$$f_{Y_i}(y_i;\theta) = \Pr(Y_i = y_i) = \theta^{y_i}(1-\theta)^{1-y_i}$$

We assume that the (uninformative) prior distribution for θ is an (continuous) uniform distribution over $[0,1]$.

Question: Write the pdf associated to the unormalised posterior distribution and the posterior distribution

2. Prior and posterior distributions

Solution

$$\pi\left(\theta \middle| y_1, ..y_n\right) = \frac{L_n\left(\theta; y_1, .., y_n\right) \times \pi\left(\theta\right)}{f_{Y_1,..Y_n}\left(y_1, .., y_n\right)}$$

The sample $\left(y_1, .., y_n\right)$ is $i.i.d.$, so its **likelihood** is defined as to be:

$$L_n\left(\theta; y_1, .., y_n\right) = \prod_{i=1}^{n} f_{Y_i}\left(y_i\right) = \prod_{i=1}^{n} \theta^{y_i}\left(1-\theta\right)^{1-y_i}$$

So, we have:

$$L_n\left(\theta; y_1 .., y_n\right) = \theta^{\Sigma y_i}\left(1-\theta\right)^{\Sigma\left(1-y_i\right)}$$

2. Prior and posterior distributions

Solution (cont'd)

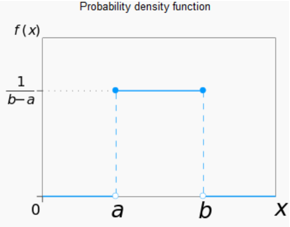
$$\pi\left(\theta \mid y_1,..y_n\right) = \frac{L_n\left(\theta; y_1,..,y_n\right) \times \pi\left(\theta\right)}{f_{Y_1,..Y_n}\left(y_1,..,y_n\right)}$$

The (uninformative) **prior distribution** is

$$\theta \sim U_{[0,1]}$$

with a pdf:

$$\pi\left(\theta\right) = \begin{cases} 1 & \text{if } \theta \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$



Source: wikipedia

Solution (cont'd)

$$\pi\left(\theta \mid y_1, ..y_n\right) \propto L_n\left(\theta; y_1, .., y_n\right) \times \pi\left(\theta\right)$$

$$L_n\left(\theta; y_1 .., y_n\right) = \theta^{\Sigma y_i}\left(1-\theta\right)^{\Sigma(1-y_i)}$$

$$\pi\left(\theta\right) = \begin{cases} 1 & \text{if } \theta \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

The **unormalised** posterior distribution is

$$L_n\left(\theta; y_1, .., y_n\right) \times \pi\left(\theta\right) = \begin{cases} \theta^{\Sigma y_i}\left(1-\theta\right)^{\Sigma(1-y_i)} & \text{if } \theta \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

2. Prior and posterior distributions

Solution (cont'd)

$$\pi\left(\theta \mid y_1, ..y_n\right) = \frac{L_n\left(\theta; y_1, .., y_n\right) \times \pi\left(\theta\right)}{f_{Y_1,..Y_n}\left(y_1, .., y_n\right)}$$

The **joint density** of $\left(Y_1, .., Y_n\right)$ evaluated at $\left(y_1, .., y_n\right)$

$$\begin{aligned} f_{Y_1,..Y_n}\left(y_1, .., y_n\right) &= \int_0^1 f_{Y_1,..Y_n|\theta}\left(y_1, .., y_n \mid \theta\right) \pi\left(\theta\right) d\theta \\ &= \int_0^1 L_n\left(\theta; y_1, .., y_n\right) \pi\left(\theta\right) d\theta \\ &= \int_0^1 \theta^{\Sigma y_i} \left(1-\theta\right)^{\Sigma(1-y_i)} d\theta \end{aligned}$$

Solution (cont'd)

$$\pi\left(\theta \mid y_1, ..y_n\right) = \frac{L_n\left(\theta; y_1, .., y_n\right) \times \pi\left(\theta\right)}{f_{Y_1,..Y_n}\left(y_1, .., y_n\right)}$$

Finally, we have

$$\pi\left(\theta \mid y_1, ..y_n\right) = \frac{\theta^{\Sigma y_i}\left(1-\theta\right)^{\Sigma\left(1-y_i\right)}}{\int_0^1 \theta^{\Sigma y_i}\left(1-\theta\right)^{\Sigma\left(1-y_i\right)} d\theta} \quad \text{if } \theta \in [0,1]$$

$$\pi\left(\theta \mid y_1, ..y_n\right) = 0 \quad \text{if } \theta \notin [0,1] \quad \Box$$

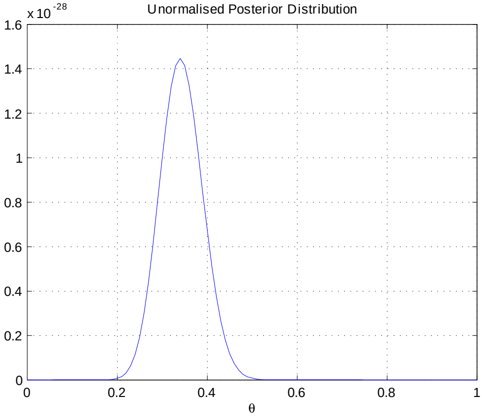
2. Prior and posterior distributions

Example (Posterior distribution)

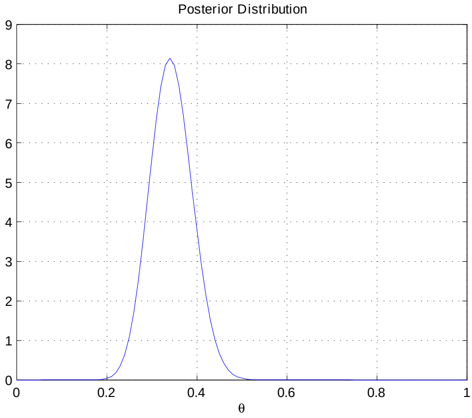
Consider an $i.i.d.$ sample $(Y_1, .., Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$ with $\theta = 0.3$. We assume that the (uninformative) prior distribution for θ is an uniform distribution over $[0, 1]$.

Question: Write a Matlab code in order (i) to generate a sample of size $n = 100$, (ii) to display the pdf associated to the prior and the posterior distribution.

2. Prior and posterior distributions



2. Prior and posterior distributions



2. Prior and posterior distributions

```
clear all ; clc ; close all
global y
theta0=0.3;                % True value for theta
n=100;                     % Sample size
y=binornd(1,theta0,n,1);   % sample of n trial in a B(1,theta)
% Prior distribution
theta=(0:0.01:1)';             % Potential values for theta
pi=(theta>=0)&(theta<=1);      % Prior distribution
% Likelihood function
Ln=theta.^(sum(y)).*(1-theta).^(n-sum(y));
% Unormalised Posterior Distribution
Un=Ln.*pi;
plot(theta,Un)
xlabel('\theta'), grid('on')
title('Unormalised Posterior Distribution')
% Posterior Distribution
fy=quad('Likelihood',0,1);
post=Ln.*pi/fy;
figure
plot(theta,post)
xlabel('\theta'),grid('on')
title('Posterior Distribution')
```

2. Prior and posterior distributions

```
function L=Likelihood(thet)
global y
n=length(y);
L=thet.^(sum(y)).*(1-thet).^(n-sum(y));
```

2. Prior and posterior distributions

Example (Posterior distribution)

Consider an *i.i.d.* sample $(Y_1, .., Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$ and:

$$f_{Y_i}(y_i;\theta) = \Pr(Y_i = y_i) = \theta^{y_i}(1-\theta)^{1-y_i}$$

We assume that the prior distribution for θ is a Beta distribution $B(\alpha,\beta)$ with a pdf:

$$\pi(\theta;\gamma) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \alpha,\beta > 0 \qquad \theta \in [0,1]$$

with $\gamma = (\alpha,\beta)^\top$ the vector of hyperparameters.

Question: Write the pdf associated to the unormalised posterior distribution.

2. Prior and posterior distributions

Solution

The **likelihood of the sample** $(y_1, .., y_n)$ is:

$$L_n(\theta; y_1.., y_n) = \theta^{\Sigma y_i} (1-\theta)^{\Sigma(1-y_i)}$$

The **prior distribution** is:

$$\pi(\theta; \gamma) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

So the **unormalised posterior distribution** is:

$$\begin{aligned}
\pi(\theta | y_1, ..y_n) &\propto L_n(\theta; y_1, .., y_n) \times \pi(\theta) \\
&\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \theta^{\Sigma y_i} (1-\theta)^{\Sigma(1-y_i)}
\end{aligned}$$

Solution (cont'd)

$$\pi\left(\theta \middle| y_1, ..y_n\right) \propto \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}\theta^{\Sigma y_i}\left(1-\theta\right)^{\Sigma\left(1-y_i\right)}$$

or equivalently:

$$\pi\left(\theta \middle| y_1, ..y_n\right) \propto \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha-1+\Sigma y_i}\left(1-\theta\right)^{\beta-1+\Sigma\left(1-y_i\right)}$$

Note that the term $\Gamma\left(\alpha+\beta\right) / \left(\Gamma\left(\alpha\right)\Gamma\left(\beta\right)\right)$ **does not depend** on θ. So, the unormalised posterior can also be written as:

$$\pi\left(\theta \middle| y_1, ..y_n\right) \propto \theta^{\left(\alpha+\Sigma y_i\right)-1}\left(1-\theta\right)^{\left(\beta+\Sigma\left(1-y_i\right)\right)-1} \quad \square$$

2. Prior and posterior distributions

Solution (cont'd)

$$\pi\left(\theta \mid y_1, ..y_n\right) \propto \theta^{(\alpha+\Sigma y_i)-1}\left(1-\theta\right)^{(\beta+\Sigma(1-y_i))-1}$$

Remind that the pdf of a Beta $B\left(\alpha, \beta\right)$ distribution is

$$\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}$$

The posterior distribution is in the form of a **beta distribution** with parameters

$$\alpha_1 = \alpha + \sum_{i=1}^{n} y_i \quad \beta_1 = \beta + n - \sum_{i=1}^{n} y_i$$

This is an example of a **conjugate prior**, where the posterior distribution is in the same family as the prior distribution

2. Prior and posterior distributions

Definition (Conjugate prior)

A **conjugate prior** is such that the posterior distribution is in the same family as the prior distribution.

Example (Posterior distribution)

Consider an $i.i.d.$ sample $(Y_1, .., Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$ and:

$$f_{Y_i}(y_i; \theta) = \Pr(Y_i = y_i) = \theta^{y_i} (1 - \theta)^{1 - y_i}$$

We assume that the prior distribution for θ is a Beta distribution $\beta(\alpha, \beta)$

Question: Determine the mean of the posterior distribution.

2. Prior and posterior distributions

Solution

We know that:

$$\pi\left(\theta \mid y_1, ..y_n\right) \propto \theta^{\alpha_1 - 1}\left(1-\theta\right)^{\beta_1 - 1}$$

$$\alpha_1 = \alpha + \sum_{i=1}^{n} y_i \qquad \beta_1 = \beta + n - \sum_{i=1}^{n} y_i$$

A simple way to normalise this posterior distribution consists in writing:

$$\pi\left(\theta \mid y_1, ..y_n\right) = \frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right)\Gamma\left(\beta_1\right)} \theta^{\alpha_1 - 1}\left(1-\theta\right)^{\beta_1 - 1}$$

This expression corresponds to the pdf of $B\left(\alpha_1, \beta_1\right)$ distribution and it is normalised to 1 by definition:

$$\int_0^1 \pi\left(\theta \mid y_1, ..y_n\right) d\theta = 1$$

2. Prior and posterior distributions

Solution (cont'd)

$$\pi\left(\theta \mid y_1, ..y_n\right) = \frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right)\Gamma\left(\beta_1\right)}\theta^{\alpha_1 - 1}\left(1-\theta\right)^{\beta_1 - 1}$$

$$\alpha_1 = \alpha + \sum_{i=1}^{n} y_i \qquad \beta_1 = \beta + n - \sum_{i=1}^{n} y_i$$

Using the properties of the $B\left(\alpha_1, \beta_1\right)$ distribution, we have:

$$\begin{aligned}
\mathbb{E}\left(\theta \mid y_1, ..y_n\right) &= \frac{\alpha_1}{\alpha_1 + \beta_1} \\
&= \frac{\alpha + \sum_{i=1}^{N} y_i}{\alpha + \sum_{i=1}^{N} y_i + \beta + n - \sum_{i=1}^{N} y_i} \\
&= \frac{\alpha + \sum_{i=1}^{N} y_i}{\alpha + \beta + n}
\end{aligned}$$

Solution (cont'd)

$$\mathbb{E}\left(\theta \mid y_1, ..y_n\right) = \frac{\alpha + \sum_{i=1}^{N} y_i}{\alpha + \beta + n}$$

We can express the mean as a function of the MLE estimator, $\overline{y}_n = n^{-1} \sum_{i=1}^{N} y_i$, as follows:

$$\underbrace{\mathbb{E}\left(\theta \mid y_1, ..y_n\right)}_{\text{posterior mean}} = \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \underbrace{\frac{\alpha}{\alpha + \beta}}_{\text{prior mean}} + \left(\frac{n}{\alpha + \beta + n}\right) \underbrace{\overline{y}_n}_{\text{MLE}} \quad \square$$

2. Prior and posterior distributions

Remark

$$\underbrace{\mathbb{E}\left(\theta \mid y_1, ..y_n\right)}_{\text{posterior mean}} = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right) \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{prior mean}} + \left(\frac{n}{\alpha+\beta+n}\right) \underbrace{\overline{y}_n}_{\text{MLE}}$$

1. If $n \longrightarrow \infty$, then the weight on the prior mean approaches zero, and the weight on the MLE approaches one, implying

$$\lim_{n\longrightarrow\infty} \mathbb{E}\left(\theta \mid y_1, ..y_n\right) = \overline{y}_n$$

2. If the sample size is very small, $n \longrightarrow 0$, then we have

$$\lim_{n\longrightarrow 0} \mathbb{E}\left(\theta \mid y_1, ..y_n\right) = \frac{\alpha}{\alpha+\beta}$$

2. Prior and posterior distributions

Example (Posterior distribution)

Consider an $i.i.d.$ sample $(Y_1, .., Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$ and

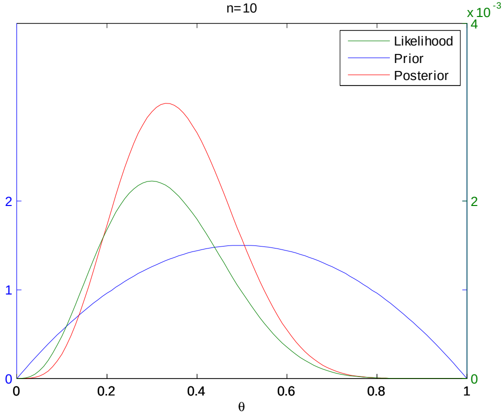
$$\sum_{i=1}^{N} y_i = 3 \text{ if } n = 10$$

$$\sum_{i=1}^{N} y_i = 15 \text{ if } n = 50$$

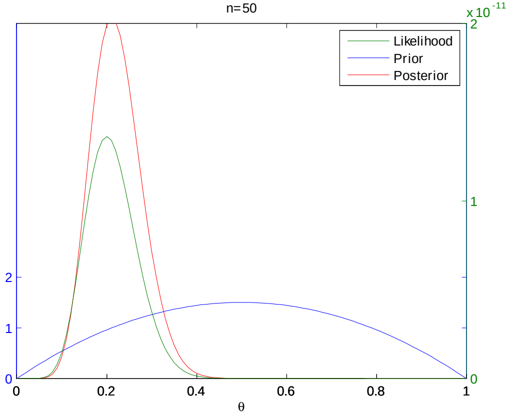
We assume that the prior distribution for θ is a Beta distribution $\beta(2, 2)$

Question: Write a Matlab code to plot (i) the prior, (ii) the posterior distribution and the (iii) the likelihood in the two cases.

2. Prior and posterior distributions



2. Prior and posterior distributions



2. Prior and posterior distributions

```
%=====================================================
% PURPOSE: Reproduce the Figure 6 of the Chapter 7
% Lecture: "Advanced Econometrics"
%-------------------------
% Author: Christophe Hurlin, University of Orleans
% Version: v1. June 2014
%=====================================================
clear all ; clc ; close all
n=50;                                   % Sample size
sumy=10;                                % Sum of y
theta=(0:0.01:1)';                      % Potential values for theta
a=2;b=2;                                % Parameters
prior=betapdf(theta,a,b);               % Prior distribution
Ln=theta.^(sumy).*(1-theta).^(n-sumy);  % Likelihood
a1=a+sumy;
b1=b+n-sumy;
post=betapdf(theta,a1,b1);              % Posterior distribution

plotyy(theta,prior,theta,Ln)
ylim([0 7])
xlabel('\theta')
hold('on')
plot(theta,post,'r')
legend('Likelihood','Prior','Posterior')
title('n=50')
```

Key concepts of Section 2

1. Frequentist versus subjective probability
2. Prior distribution
3. Hyperparameters
4. Prior predictive distribution
5. Posterior distribution
6. Unormalised posterior distribution
7. Conjugate prior

Section 3

Posterior distributions and inference

3. Posterior distributions and inference

Objectives

The objective of this section are the following:

1. Generalise the bayesian approach to a **vector of parameters**
2. Generalise the bayesian approach to a **regression model**
3. Introduce the **Bayesian updating** mechanism
4. Study the posterior distribution in the case of a **large sample**
5. Discuss the **inference** in a Bayesian framework

3. Posterior distributions and inference

The concept of posterior distribution can be generalized to:

1. a case with a **vector of parameters** $\boldsymbol{\theta} = (\theta_1, .., \theta_d)^\top$.
2. a **model** with exogenous variable and/or lagged endogenous variables (linear regression model, time series models, DSGE etc.).

Sub-Section 3.1

Generalisation to a vector of parameters

Vector of parameters

Consider a model/variable with a pdf/pmf that depends on a **vector of parameters** $\boldsymbol{\theta} = (\theta_1, .., \theta_d)^\top$.

- The previous definitions of likelihood, prior, and posterior distributions still apply.
- But they are now, respectively, the **joint** prior distribution and **joint** posterior distribution of the **multivariate random variable** $\boldsymbol{\theta}$
- From the joint distributions, we may derive **marginal** and **conditional** distributions.

Definition (Marginal posterior distribution)

The **marginal posterior distribution** of $\theta_1$can be found by integrating out the remainder of the parameters from the joint posterior distribution:

$$\pi\left(\theta_1 \middle| y\right) = \int \pi\left(\theta_1, .., \theta_d \middle| y\right) d\theta_2..d\theta_d$$

Definition (Conditional posterior distribution)

The **conditional posterior distribution** of $\theta_1$is defined as to be:

$$\pi\left(\theta_1 \mid \theta_2, .., \theta_d, y\right) = \frac{\pi\left(\theta_1, .., \theta_d \mid y\right)}{\pi\left(\theta_2, .., \theta_d \mid y\right)}$$

where the denominator on the right-hand side is the marginal posterior distribution of $\left(\theta_2, .., \theta_d\right)$ obtained by integrating θ_1 from the joint distribution.

Remark

In most applications, the **marginal distribution** of a parameter is more useful than its **conditional distribution** because the marginal takes into account the uncertainty over the values of the remaining parameters, while the conditional sets them at particular values.

Example (Marginal posterior distribution)

Consider the **multinomial distribution** $Mn\,(.)$, which generalizes the Bernoulli example discussed above. In this model, each trial, assumed independent of the other trials, results in one of d outcomes, labeled $1, 2,..,d$ with probabilities $\theta_1, .., \theta_d$ where $\sum_{i=1}^{d} \theta_i = 1$. When the experiment is repeated n times and outcome i arises y_i times, the likelihood function is

$$L_n\left(\boldsymbol{\theta} \mid\, y_1, .., y_n\right) = \theta_1^{y_1} \theta_2^{y_2} ... \theta_d^{y_d} \quad \text{with } \sum_{i=1}^{d} y_i = n$$

Example (cont'd)

Consider a **Dirichlet** prior distribution (generalisation of the Beta) with:

$$\pi(\boldsymbol{\theta}) = \frac{\Gamma\left(\sum_{i=1}^{d} \alpha_i\right)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_d^{\alpha_d - 1}, \qquad \alpha_i > 0, \ \sum_{i=1}^{d} \theta_i = 1$$

Question: Determine the marginal posterior distribution of θ_1.

Solution

Following the previous procedure, we find the (unormalised) posterior joint distribution given the data $y = (y_1, .., y_n)$:

$$\begin{aligned} \pi(\boldsymbol{\theta} | y_1, ..y_n) &\propto L_n(\boldsymbol{\theta} | y_1, .., y_n) \times \pi(\boldsymbol{\theta}) \\ &\propto \theta_1^{y_1} \theta_2^{y_2} ... \theta_d^{y_d} \frac{\Gamma\left(\sum_{i=1}^{d} \alpha_i\right)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} ... \theta_d^{\alpha_d - 1} \end{aligned}$$

Or equivalently:

$$\pi(\boldsymbol{\theta} | y_1, ..y_n) \propto \theta_1^{y_1 + \alpha_1 - 1} \theta_2^{y_2 + \alpha_2 - 1} ... \theta_d^{y_d + \alpha_d - 1}$$

Remark: the Dirichlet prior is a **conjugate** prior for the multinomial model.

Solution (cont'd)

The marginal distribution of θ_1 is defined as to be:

$$\pi\left(\theta_1 \mid y_1,..,y_n\right) = \int_0^1 \pi\left(\boldsymbol{\theta} \mid y_1,..y_n\right) d\theta_2..d\theta_d$$

In this context, we have:

$$\begin{aligned} & \pi\left(\theta_1 \mid y_1,..,y_n\right) \\ = \quad & \frac{\Gamma\left(\sum_{i=1}^{p} \alpha_i + y_i\right)}{\prod_{i=1}^{p} \Gamma\left(\alpha_i + y_i\right)} \theta_1^{y_1+\alpha_1-1} \int_0^1 \theta_2^{y_2+\alpha_2-1} ... \theta_d^{y_d+\alpha_d-1} d\theta_2..d\theta_d \end{aligned}$$

But, we can also use some general results about the Dirichlet distribution.

Solution (cont'd)

Definition (Dirichlet distribution)

The Dirichlet distribution generalizes the Beta distribution. Let $x = (x_1, .., x_p)$ with $0 \leq x_i \leq 1$, $\sum_{i=1}^{p} x_i = 1$. Then $x \sim D(\alpha_1, .., \alpha_p)$ if

$$f(x; \alpha_1, .\alpha_p) = \frac{\Gamma\left(\sum_{i=1}^{p} \alpha_i\right)}{\prod_{i=1}^{p} \Gamma(\alpha_i)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} ... x_d^{\alpha_d - 1}, \quad \alpha_i > 0$$

Marginally, we have

$$x_i \sim B\left(\alpha_i, \sum_{k \neq i} \alpha_k\right)$$

Solution (cont'd)

$$\pi\left(\boldsymbol{\theta}\right| y_1, ..y_n) \propto \theta_1^{y_1+\alpha_1-1}\theta_2^{y_2+\alpha_2-1}...\theta_d^{y_d+\alpha_d-1}$$

$$\boldsymbol{\theta}| \, y_1, ..y_n \sim D\left(y_1+\alpha_1, .., y_p+\alpha_p\right)$$

The marginal posterior distribution of θ_1 is a Beta distribution:

$$\theta_1| \, y_1, ..y_n \sim B\left(y_1+\alpha_1, \textstyle\sum_{i=2}^{p} y_i+\alpha_i\right) \quad \square$$

Sub-Section 3.2

Generalisation to a model

Remark: We can also aim at estimating the parameters of a **model (with dependent and explicative variables)** such that:

$$y = g(x; \theta) + \varepsilon$$

where θ denotes the vector or parameters, X a set of explicative variables, ε and error term and $g(.)$ the link function.

In this case, we generally consider the *conditional distribution* of Y given X, which is equivalent to unconditional distribution of the error term ε :

$$Y|\,X \sim D \Longleftrightarrow \varepsilon \sim D$$

Notations (model)

- Let us consider two continuous random variables Y and X
- We assume that Y has a conditional distribution given $X = x$ with a pdf denoted $f_{Y|x}(y;\theta)$, for $y \in \mathbb{R}$
- $\theta = (\theta_1..\theta_K)^{\intercal}$ is a $K \times 1$ vector of unknown parameters. We assume that $\theta \in \Theta \subset \mathbb{R}^K$.
- Let us consider a sample $\{X_1, Y_N\}_{i=1}^{n}$ of $i.i.d.$ random variables and a realisation $\{x_1, y_N\}_{i=1}^{n}$.

Definition (Conditional likelihood function)

The (conditional) likelihood function of the $i.i.d.$ sample $\{X_i, Y_i\}_{i=1}^{n}$ is defined to be:

$$L_n(\theta; y|x) = \prod_{i=1}^{n} f_{Y|X}(y_i|x_i;\theta)$$

where $f_{Y|X}(y_i|x_i;\theta)$ denotes the conditional pdf of Y_i given X_i.

Remark: The conditional likelihood function is the joint conditional density of the data given θ.

Example (Linear Regression Model)

Consider the following linear regression model:

$$y_i = \mathbf{X}_i^{\top} \boldsymbol{\beta} + \varepsilon_i$$

where $\mathbf{X}_i$ is a $K \times 1$ vector of random variables and $\boldsymbol{\beta} = (\beta_1 .. \beta_K)^{\top}$ a $K \times 1$ vector of parameters. We assume that the ε_i are $i.i.d.$ with $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$. Then, the conditional distribution of Y_i given $\mathbf{X}_i = \mathbf{x}_i$ is:

$$Y_i | \, \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{x}_i^{\top} \boldsymbol{\beta}, \sigma^2\right)$$

$$L_i\left(\boldsymbol{\theta}; y | \, \mathbf{x}\right) = f_{Y|\mathbf{x}}\left(y_i | \, \mathbf{x}_i; \boldsymbol{\theta}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^{\top} \boldsymbol{\beta}\right)^2}{2\sigma^2}\right)$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^{\top} \; \sigma^2\right)^{\top}$ is $K + 1 \times 1$ vector.

Example (Linear Regression Model, cont'd)

Then, if we consider an $i.i.d.$ sample $\{y_i, \mathbf{x}_i\}_{i=1}^n$, the corresponding **conditional** (log-)likelihood is defined to be:

$$\begin{aligned} L_n(\boldsymbol{\theta}; y|\mathbf{x}) &= \prod_{i=1}^n f_{Y|X}(y_i|\mathbf{x}_i; \boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{2\sigma^2}\right) \\ &= \left(\sigma^2 2\pi\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2\right) \end{aligned}$$

$$\ell_n(\boldsymbol{\theta}; y|\mathbf{x}) = -\frac{n}{2}\ln\left(\sigma^2\right) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2$$

Remark: Given this principle, we can derive the (conditional) likelihood and the log-likelihood functions associated to a specific sample for any type of econometric model in which the conditional distribution of the dependent variable is known.

- Dichotomic models: probit, logit models etc.
- Censored regression models: Tobit etc.
- Times series models: AR, ARMA, VAR etc.
- GARCH models
- ....

Definition (Posterior distribution, model)

For the sample $\{y_i, x_i\}_{i=1}^{n}$, the **posterior distribution** is the conditional distribution of θ given y_i, defined as to be:

$$\pi(\theta | y, x) = \frac{L_n(\theta; y | x) \times \pi(\theta)}{f_{Y|X}(y | x)}$$

where $L_n(\theta; y | x)$ is the likelihood of the sample and

$$f_{Y|X}(y | x) = \int_{\Theta} L_n(\theta; y | x) \pi(\theta) d\theta$$

and Θ the support of the distribution of θ.

Sub-Section 3.3

Bayesian updating

Bayesian updating

A very attractive feature of Bayesian inference is the way in which posterior distributions are **updated** as new information becomes available.

Bayesian updating

As usual for the first observation y_1, we have

$$\pi\left(\theta \mid y_1\right) \propto f\left(y_1 \mid \theta\right) \pi\left(\theta\right)$$

Next, suppose that a new data y_2 is obtained, and we wish to compute the posterior distribution given the complete data set $\pi\left(\theta \mid y_1, y_2\right)$:

$$\pi\left(\theta \mid y_1, y_2\right) \propto f\left(y_1, y_2 \mid \theta\right) \pi\left(\theta\right)$$

The posterior can be rewritten as:

$$\begin{aligned} \pi\left(\theta \mid y_1, y_2\right) & \propto f\left(y_2 \mid y_1, \theta\right) f\left(y_1 \mid \theta\right) \pi\left(\theta\right) \\ & \propto f\left(y_2 \mid y_1, \theta\right) \pi\left(\theta \mid y_1\right) \end{aligned}$$

Definition (Bayesian updating)

The posterior distribution is **updated** as new information becomes available as follows:

$$\pi\left(\theta \mid y_1, .., y_n\right) \propto f\left(y_n \mid y_{n-1}, \theta\right) \pi\left(\theta \mid y_1, .., y_{n-1}\right)$$

If the observations y_n and y_{n-1} are independent ($i.i.d.$ sample), then $f\left(y_n \mid y_{n-1}, \theta\right) = f\left(y_n \mid \theta\right)$ and

$$\pi\left(\theta \mid y_1, .., y_n\right) \propto f\left(y_n \mid \theta\right) \pi\left(\theta \mid y_1, .., y_{n-1}\right)$$

3. Posterior distributions and inference

3.3. Bayesian updating

Example (Bayesian updating)

Consider an $i.i.d.$ sample $(Y_1, .., Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$ and:

$$f_{Y_i}(y_i; \theta) = \Pr(Y_i = y_i) = \theta^{y_i} (1-\theta)^{1-y_i}$$

We assume that the prior distribution for θ is a Beta distribution $\beta(\alpha, \beta)$ with a pdf:

$$\pi(\theta; \gamma) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \qquad \alpha, \beta > 0 \qquad \theta \in [0,1]$$

with $\gamma = (\alpha, \beta)^{\top}$ the vector of hyperparameters.

Question: Write the posterior of θ given (y_1, y_2) as a function of the posterior given y_1.

Solution

The **likelihood of** (y_1, y_2) is:

$$L(\theta; y_1, y_2) = \theta^{y_1+y_2}(1-\theta)^{2-y_1-y_2}$$

The **prior distribution** is:

$$\pi(\theta; \gamma) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

So the **unormalised posterior distribution** is:

$$\begin{aligned}\pi(\theta | y_1, y_2) &\propto L(\theta; y_1, y_2) \times \pi(\theta) \\ &\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{y_1+y_2}(1-\theta)^{2-y_1-y_2}\end{aligned}$$

Solution (cont'd)

$$\pi(\theta|\, y_1, y_2) \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{y_1+y_2}(1-\theta)^{2-y_1-y_2}$$

or equivalently:

$$\pi(\theta|\, y_1, y_2) \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha+y_1+y_2-1}(1-\theta)^{\beta+2-y_1-y_2-1}$$

So,

$$\theta|\, y_1, y_2 \sim B(\alpha+y_1+y_2, \beta+2-y_1-y_2)$$

Solution (cont'd)

$$\pi\left(\theta \mid y_1, y_2\right) \propto \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha+y_1+y_2-1}\left(1-\theta\right)^{\beta+2-y_1-y_2-1}$$

$$\theta \mid y_1, y_2 \sim B\left(\alpha+y_1+y_2, \beta+2-y_1-y_2\right)$$

Given the observation y_1, we have:

$$\pi\left(\theta \mid y_1\right) \propto \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha+y_1-1}\left(1-\theta\right)^{\beta+1-y_1-1}$$

$$\theta \mid y_1 \sim B\left(\alpha+y_1, \beta+1-y_1\right)$$

Solution (cont'd)

$$\pi\left(\theta | y_1, y_2\right) \propto \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha+y_1+y_2-1}\left(1-\theta\right)^{\beta+2-y_1-y_2-1}$$

$$\pi\left(\theta | y_1\right) \propto \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha+y_1-1}\left(1-\theta\right)^{\beta+1-y_1-1}$$

The updating mechanism is given by:

$$\pi\left(\theta | y_1, y_2\right) \propto \pi\left(\theta | y_1\right)\ \theta^{y_2}\left(1-\theta\right)^{1-y_2}$$

or equivalently

$$\pi\left(\theta | y_1, y_2\right) \propto \pi\left(\theta | y_1\right)\ f_{Y_2}\left(y_2;\theta\right) \quad \square$$

Sub-Section 3.4

Large sample

Large samples

Although all statistical results for Bayesian estimators are necessarily **"finite sample"** (they are conditioned on the sample data), it remains of interest to consider how the estimators behave in **large samples**.

What is the the behavior of the posterior distribution when n is large?

$$\pi\left(\theta \mid y_1, .., y_n\right)$$

Fact (Large sample)

Greenberg (2008) summarises the behavior of the posterior distribution when n is **large** *as follows:*

(1) the prior distribution plays a relatively small role in determining the posterior distribution,

(2) the posterior distribution converges to a degenerate distribution at the true value of the parameter,

(3) the posterior distribution is approximately normally distributed with mean $\widehat{\theta}$, the MLE of θ.

Large samples

What is the intuition of the two first results?

1. the prior distribution plays a relatively small role in determining the posterior distribution
2. the posterior distribution converges to a degenerate distribution at the true value of the parameter

Definition (Log-Likelihood Function)

The log-likelihood function is defined to be:

$$\ell_n : \Theta \times \mathbb{R}^n \longrightarrow \mathbb{R}$$

$$(\theta; y_1, .., y_n) \longmapsto \ell_n(\theta; y_1, .., y_n) = \ln(L_n(\theta; y_1, .., y_n)) = \sum_{i=1}^{n} \ln f_Y(y_i; \theta)$$

Large samples

Introduce the mean log likelihood contribution (cf. chapter 3):

$$\ell_n(\theta; y) = \sum_{i=1}^{n} \ln f_Y(y_i; \theta) = n\bar{\ell}(\theta; y)$$

The posterior distribution can be written as

$$\pi(\theta | y) \propto \underbrace{\exp\left(n\bar{\ell}(\theta; y)\right)}_{\text{depends on } n} \quad \underbrace{\pi(\theta)}_{\text{does not depend on } n}$$

Large samples

$$\pi\left(\theta \mid y\right) \propto \underbrace{\exp\left(n\bar{\ell}\left(\theta; y\right)\right)}_{\text{depends on } n} \underbrace{\pi\left(\theta\right)}_{\text{does not depend on } n}$$

For large n, the exponential term dominates $\pi\left(\theta\right)$:

- The prior distribution will play a relatively smaller role than do the data (likelihood function), when the sample size is large.
- Conversely, the prior distribution has relatively greater weight when n is small

If we denote the true value of θ by θ_0, it can be shown that

$$\lim_{n \to \infty} n\bar{\ell}\left(\theta; y\right) = n\bar{\ell}\left(\theta_0; y\right)$$

Then, we have for n large:

$$\pi\left(\theta | y\right) \propto \underbrace{\exp\left(n\bar{\ell}\left(\theta_0; y\right)\right)}_{\text{does not depend on } \theta} \pi\left(\theta\right) \quad \forall \theta \in \Theta$$

Whatever the value of θ, the value of $\pi\left(\theta | y\right)$ tends to a constant times the prior.

Example (Large sample)

Consider an $i.i.d.$ sample $(Y_1, .., Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$ with $\theta = 0.3$ and:

$$f_{Y_i}(y_i; \theta) = \Pr(Y_i = y_i) = \theta^{y_i} (1-\theta)^{1-y_i}$$

We assume that the prior distribution for θ is a Beta distribution $B(2, 2)$.

Question: Write a Matlab code to illustrate the changes in the posterior distribution with n.

3. Posterior distributions and inference

3.4. Large sample

An animation is worth 1,000,000 words...

Click me!

```
n=0;y=[];                                    % Initialisation
theta=(0:0.01:1)';                           % Potential values for theta
a=2;b=2;                                     % HyperParameters
prior=betapdf(theta,a,b);                    % Prior distribution
aviobj = avifile('Chapter7_Movie1.avi','fps',2,'compression','None');
for i=1:101
    pas=10;                                   % Increment for the sample size
    n=n+pas;                                  % Sample size
    theta0=0.3;                               % True value
    y=[y ; binornd(1,theta0,pas,1)];          % We keep the previous trials
                                              % and increase the sample with
                                              % new trials
    post=betapdf(theta,a+sum(y),b+n-sum(y)); % Posterior distribution

    h=figure;
    plot(theta,prior,'Linewidth',1.5,'Color','b')
    hold('on')
    plot(theta,post,'Linewidth',1.5,'Color','r')
    ylim([0 30])
    legend('Prior','Posterior')
    movegui(h)
    F(i)=getframe(h);
    aviobj = addframe(aviobj,F(i));
    close all
end
aviobj = close(aviobj);
```

Sub-Section 3.5

Inference

The **output** of the Bayesian estimation is the posterior distribution

$$\pi\left(\theta \mid y\right)$$

1. In some particular case, the posterior distribution is a **standard distribution** (Beta, normal etc..) and its pdf has an analytic form.
2. In other case, the posterior distribution is get from **numerical simulations** (cf. section 5).

Whatever the case (analytical or numerical), the outcome of the Bayesian estimation procedure may be:

1. The **graph** of the posterior density $\pi\left(\theta | y\right)$ for all values of $\theta \in \Theta$
2. Some particular **moments** (expectation, variance etc..) or **quantiles** of this distribution

3. Posterior distributions and inference

3.5. Inference

TABLE 2A

PRIOR AND POSTERIOR DISTRIBUTION OF THE STRUCTURAL PARAMETERS

	Parameter	Prior distribution			Posterior distribution			
		Distrib.	Mean	Std.dev.	Mean	2.5%	Median	97.5%
κ_p	p stickiness	Gamma	50.0	20.0	30.57	10.68	28.65	49.89
κ_w	w stickiness	Gamma	50.0	20.0	102.35	70.29	99.90	133.81
κ_i	Invest. adj. cost	Gamma	2.5	1.0	10.26	7.57	10.18	12.81
κ_d	Dep. rate adj. cost	Gamma	10.0	2.5	3.63	2.28	3.50	4.96
κ_{bE}	Firms rate adj. cost	Gamma	3.0	2.5	9.51	6.60	9.36	12.31
κ_{bH}	HHs rate adj. cost	Gamma	6.0	2.5	10.22	7.47	10.09	12.88
κ_{Kb}	Leverage dev. cost	Gamma	10.0	5.0	11.49	4.03	11.07	18.27
ϕ_π	T.R. coeff. on π	Gamma	2.0	0.5	2.01	1.72	1.98	2.30
ϕ_R	T.R. coeff. on R	Beta	0.75	0.10	0.77	0.72	0.77	0.81
ϕ_y	T.R. coeff. on y	Normal	0.10	0.15	0.35	0.15	0.35	0.55
ι_p	p indexation	Beta	0.50	0.15	0.17	0.06	0.16	0.28
ι_w	w indexation	Beta	0.50	0.15	0.28	0.16	0.28	0.39
a^h	Habit coefficient	Beta	0.50	0.10	0.85	0.81	0.86	0.90

NOTE: Results based on 10 chains, each with 100,000 draws based on the Metropolis algorithm.

Source: Gerali et al. (2014), Credit and Banking in a DSGE Model of the Euro Area, JMCB, 42(6) 107-141.

3.5. Inference

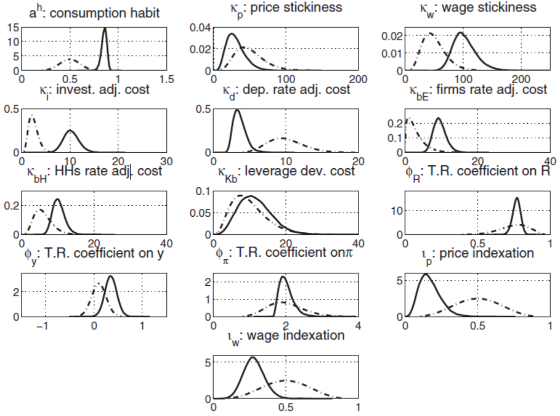


FIG. 3. Prior and Posterior Marginal Distributions.

NOTE: The marginal posterior densities are based on 10 chains, each with 100,000 draws based on the Metropolis algorithm. Solid lines denote the posterior distribution, dashed lines the prior distribution.

Source: Gerali et al. (2014), Credit and Banking in a DSGE Model of the Euro Area, JMCB, 42(6) 107-141.

But, we may also be interested in **estimating** one parameter of the model.

The Bayesian approach to this problem uses the idea of a **loss function**

$$L\left(\widehat{\theta},\theta\right)$$

where $\widehat{\theta}$ is the Bayesian estimator (cf. chapter 2).

Example (Absolute loss function)

The absolute loss function is defined to be:

$$L\left(\widehat{\theta},\theta\right)=\left|\widehat{\theta}-\theta\right|$$

Example (Quadratique loss function)

The quadratic loss function is defined to be:

$$L\left(\widehat{\theta},\theta\right)=\left(\widehat{\theta}-\theta\right)^{2}$$

Definition (Bayes estimator)

The Bayes estimator of θ is the value of $\widehat{\theta}$ that minimizes the expected value of the loss, where the expectation is taken over the posterior distribution of θ; that is, $\widehat{\theta}$ is chosen to minimize

$$\mathbb{E}\left(L\left(\widehat{\theta},\theta\right)\right)=\int L\left(\widehat{\theta},\theta\right)\pi\left(\theta\right|y)\,d\theta$$

Idea

The idea is to minimise the average loss whatever the possible value of θ

$$\mathbb{E}\left(L\left(\widehat{\theta},\theta\right)\right)=\int L\left(\widehat{\theta},\theta\right)\pi\left(\theta|\,y\right)d\theta$$

Under **quadratic loss**, we minimize

$$\mathbb{E}\left(L\left(\widehat{\theta},\theta\right)\right)=\int\left(\widehat{\theta}-\theta\right)^{2}\pi\left(\theta\right|y)\,d\theta$$

By differentiating the function with respect to $\widehat{\theta}$ and setting the derivative equal to zero

$$2\int\left(\widehat{\theta}-\theta\right)\pi\left(\theta\right|y)\,d\theta=0$$

or equivalently

$$\widehat{\theta}=\int\theta\;\pi\left(\theta\right|y)\;\;d\theta=\mathbb{E}\left(\theta\right|y)$$

Definition (Bayes estimator, quadratic loss)

For a quadratic loss function, the optimal Bayes estimator is the expectation of the posterior distribution

$$\widehat{\theta} = \mathbb{E}\left(\theta \mid y\right) = \int \theta \; \pi\left(\theta \mid y\right) \; d\theta$$

TABLE 16.2 Probit Estimates for Grade Equation

	Maximum Likelihood		*Posterior Means and Std. Devs*	
Variable	*Estimate*	*Standard Error*	*Posterior Mean*	*Posterior S.D.*
Constant	−7.4523	2.5425	−8.6286	2.7995
GPA	1.6258	0.6939	1.8754	0.7668
TUCE	0.05173	0.08389	0.06277	0.08695
PSI	1.4263	0.5950	1.6072	0.6257

Source: Greene W. (2007), Econometric Analysis, sixth edition, Pearson - Prentice Hil

In addition to reporting a point estimate of a parameter θ, it is often useful to report an interval estimate of the form

$$\Pr\left(\theta_L \leq \theta \leq \theta_U\right) = 1 - \alpha$$

Bayesians call such intervals **credibility intervals** (or Bayesian confidence intervals) to distinguish them from a quite different concept that appears in frequentist statistics, the confidence interval (cf. chapter 2).

Definition (Bayesian confidence interval)

If the posterior distribution is unimodal, then a Bayesian confidence interval or credibility interval on the value of θ, is given by the two values θ_L and θ_U such that:

$$\Pr\left(\theta_L \leq \theta \leq \theta_U\right) = 1 - \alpha$$

where α is the level of risk.

For a Bayesian, values of θ_L and θ_L can be determined to obtain the desired probability from the posterior distribution.

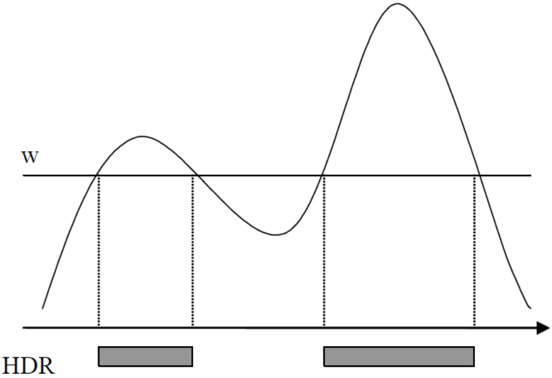
If more than one pair is possible, the pair that results in the shortest interval may be chosen; such a pair yields the **highest posterior density interval** (h.p.d.).

Definition (Highest posterior density interval)

The highest posterior density interval (hpd) is the smallest region H such that

$$\Pr(\theta \in H) = 1 - \alpha$$

where α is the level of risk. If the posterior distribution is multimodal, the hpd may be disjoint.



Source: Colletaz et Hurlin (2005), Modèles non linéaires et prévision, Rapport pour l'Institut pour la Recherche CDC.

Another basic issue in statistical inference is the **prediction** of new data values.

Definition (Forecasting)

The general form of the pdf/pmf of Y_{n+1} given $y_1, .., y_n$ is:

$$f\left(y_{n+1} \mid y\right) = \int f\left(y_{n+1} \mid \theta, y\right) \pi\left(\theta \mid y\right) d\theta$$

where $\pi\left(\theta \mid y\right)$ is the posterior distribution of θ.

Forecasting

$$f\left(y_{n+1}\mid y\right)=\int f\left(y_{n+1}\mid\theta,y\right)\pi\left(\theta\mid y\right)d\theta$$

1. If Y is a **discrete** variable, this formula gives the conditional probability of $Y_{n+1}=y_{n+1}$ given $y_1,..,y_n$:

$$\Pr\left(Y_{n+1}=y_{n+1}\mid y\right)=\int p\left(y_{n+1}\mid\theta,y\right)\pi\left(\theta\mid y\right)d\theta$$

2. If Y is a **continuous** variable, this formula gives the conditional denisty of Y_{n+1} given $y_1,..,y_n$. In this case, we can compute the expected value of Y_{n+1} as:

$$\mathbb{E}\left(y_{n+1}\mid y\right)=\int f\left(y_{n+1}\mid y\right)y_{n+1}dy_{n+1}$$

Forecasting

$$f\left(y_{n+1}\right|y) = \int f\left(y_{n+1}\right|\theta, y)\,\pi\left(\theta\right|y)\,d\theta$$

If Y_{n+1} and $Y_1, .., Y_n$ are **independent**, then $f\left(y_{n+1}\right|\theta)$ and we have:

$$f\left(y_{n+1}\right|y) = \int f\left(y_{n+1}\right|\theta)\,\pi\left(\theta\right|y)\,d\theta$$

Example (Forecasting)

Consider an $i.i.d.$ sample $(Y_1, .., Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$. We assume that the prior distribution for θ is a Beta distribution $B(\alpha, \beta)$. We want to forecast the value of Y_{n+1} given the realisations $(y_1, .., y_n)$.

Question: Determine the probability $\Pr(Y_{n+1} = 1 | y_1, .., y_n)$.

Solution

- In this example, the trials are independent, so Y_{n+1} is independent of $(Y_1, .., Y_n)$.

$$\Pr\left(Y_{n+1} = 1 \middle| y\right) = \int \Pr\left(Y_{n+1} = 1 \middle| \theta, y\right) \pi\left(\theta \middle| y\right) d\theta = \int \Pr\left(Y_{n+1} =\right.$$

- The posterior distribution of θ is is in the form of a beta distribution:

$$\pi\left(\theta \middle| y\right) = \frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right) + \Gamma\left(\beta_1\right)} \theta^{\alpha_1 - 1} \left(1 - \theta\right)^{\beta_1 - 1}$$

with

$$\alpha_1 = \alpha + \sum_{i=1}^{n} y_i \qquad \beta_1 = \alpha + n - \sum_{i=1}^{n} y_i$$

Solution (cont'd)

$$\Pr\left(Y_{n+1} = 1 \middle| y \right) = \int \Pr\left(Y_{n+1} = 1 \middle| \theta \right) \pi\left(\theta \middle| y \right) d\theta$$

$$\pi\left(\theta \middle| y \right) = \frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right) + \Gamma\left(\beta_1\right)} \theta^{\alpha_1 - 1} \left(1 - \theta\right)^{\beta_1 - 1}$$

Since $\Pr\left(Y_{n+1} = 1 \middle| \theta \right) = \theta$, we have

$$\begin{aligned} \Pr\left(Y_{n+1} = 1 \middle| y \right) &= \frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right) + \Gamma\left(\beta_1\right)} \times \int_0^1 \theta \; \theta^{\alpha_1 - 1} \left(1 - \theta\right)^{\beta_1 - 1} d\theta \\ &= \frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right) + \Gamma\left(\beta_1\right)} \times \int_0^1 \theta^{\alpha_1} \left(1 - \theta\right)^{\beta_1 - 1} d\theta \end{aligned}$$

Solution (cont'd)

$$\Pr\left(Y_{n+1}=1|\,y\right)=\frac{\Gamma\left(\alpha_1+\beta_1\right)}{\Gamma\left(\alpha_1\right)+\Gamma\left(\beta_1\right)}\times\int_0^1\theta^{\alpha_1}\left(1-\theta\right)^{\beta_1-1}d\theta$$

We admit that:

$$\int_0^1\theta^{\alpha_1}\left(1-\theta\right)^{\beta_1-1}d\theta=\frac{\Gamma\left(\alpha_1+1\right)\Gamma\left(\beta_1\right)}{\Gamma\left(\alpha_1+\beta_1+1\right)}$$

So, we have

$$\Pr\left(Y_{n+1}=1|\,y\right)=\frac{\Gamma\left(\alpha_1+\beta_1\right)}{\Gamma\left(\alpha_1\right)+\Gamma\left(\beta_1\right)}\times\frac{\Gamma\left(\alpha_1+1\right)\Gamma\left(\beta_1\right)}{\Gamma\left(\alpha_1+\beta_1+1\right)}$$

Solution (cont'd)

$$\Pr\left(Y_{n+1} = 1 \right| y) = \frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right) + \Gamma\left(\beta_1\right)} \times \frac{\Gamma\left(\alpha_1 + 1\right)\Gamma\left(\beta_1\right)}{\Gamma\left(\alpha_1 + \beta_1 + 1\right)}$$

Since $\Gamma\left(p\right) = \left(p-1\right)\,\Gamma\left(p-1\right)$, we have:

$$\begin{aligned} \Pr\left(Y_{n+1} = 1 \right| y) &= \frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right) + \Gamma\left(\beta_1\right)} \times \frac{\Gamma\left(\alpha_1\right)\ \alpha_1\ \Gamma\left(\beta_1\right)}{\Gamma\left(\alpha_1 + \beta_1\right)\left(\alpha_1 + \beta_1\right)} \\ &= \frac{\alpha_1}{\alpha_1 + \beta_1} \end{aligned}$$

Solution (cont'd)

$$\Pr\left(Y_{n+1}=1\right|y)=\frac{\alpha_1}{\alpha_1+\beta_1}=\frac{\alpha+\sum_{i=1}^{n}y_i}{\alpha+\beta+n}$$

So, we found that

$$\Pr\left(Y_{n+1}=1\right|y)=\mathbb{E}\left(\theta\right|y)=\frac{\alpha+\sum_{i=1}^{n}y_i}{\alpha+\beta+n}\quad\square$$

The estimate of $\Pr\left(Y_{n+1}=1\right|y)$ is the mean of the posterior distribution of θ.

3. Posterior distributions and inference

Key concepts of Section 3

1. Marginal and conditional posterior distribution
2. Bayesian updating
3. Bayes estimator
4. Bayesian confidence interval
5. Highest posterior density interval
6. Bayesian prediction

Section 4

Applications

5. Applications

Objectives

The objective of this section are the following:

1. To discuss the Bayesian estimation of **VAR** models
2. To propose various **priors** for this type of models
3. To introduce the issue of **numerical simulations** of the posterior

5. Applications

Consider a typical VAR(p)

$$\mathbf{Y}_t = \mathbf{B}_1 \mathbf{Y}_{t-1} + \mathbf{B}_2 \mathbf{Y}_{t-2} + .. + \mathbf{B}_p \mathbf{Y}_{t-p} + \mathbf{D}\mathbf{z}_t + \varepsilon_t \qquad t = 1, .., T$$

- $\mathbf{Y}_t$ is $n \times 1$ vector of endogenous variables
- ε_t is a $n \times 1$ vector of error terms $i.i.d.$ with

$$\varepsilon_t \sim IIN\left(\underset{n\times 1}{\mathbf{0}}, \underset{n\times n}{\boldsymbol{\Sigma}}\right)$$

- $\mathbf{z}_t$ is a $d \times 1$ vector of exogenous variables
- $\mathbf{B}_i$ for $i = 1, .., p$ is a $n \times n$ matrix of parameters
- $\mathbf{D}$ is a $n \times d$ matrix of parameters

5. Applications

- Classical estimation of the parameters $\mathbf{B}_1, .., \mathbf{B}_p, \mathbf{D}, \mathbf{\Sigma}$ may yield imprecisely estimated relations that fit the data well only because of the large number of variables included: problem known as **overfitting**
- The number of parameters to be estimated $n\left(np + d + (n-1)/2\right)$ grows **geometrically** with the number of variables (n) and **proportionally** with the number of lags (p).
- When the number of parameters is relatively high and the sample information is relatively loose (macro-data), it is likely that the estimates are influenced by noise as opposed to signal

5. Applications

A Bayesian approach to VAR estimation was originally advocated by Litterman (1980) as solution to the **overfitting** problem.

Litterman R. (1980), Techniques for forecasting with Vector AutoRegressions, University of Minnesota, Ph.D. dissertation.

5. Applications

- Bayesian estimation is a solution to the overfitting that avoids imposing exact zero restrictions on the parameters
- The researcher cannot be sure that some coefficients are zero and should not ignore their possible range of variations
- A Bayesian perspective fits precisely this view with the **prior distribution**.

5. Applications

$$\mathbf{Y}_t = \mathbf{B}_1 \mathbf{Y}_{t-1} + \mathbf{B}_2 \mathbf{Y}_{t-2} + .. + \mathbf{B}_p \mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t$$

Rewrite the VAR in a compact form:

$$\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t$$

- $\mathbf{X}_t = I_n \otimes W_{t-1}$ is $n \times nk$ with $k = p + d$
- $W_{t-1} = \left(\mathbf{Y}_{t-1}^{\top}, .., \mathbf{Y}_{t-p}^{\top}, \mathbf{z}_t^{\top} \right)^{\top}$ is $k \times 1$
- $\boldsymbol{\beta} = vec\left(\mathbf{B}_1, .., \mathbf{B}_p, \mathbf{D} \right)$ is a $nk \times 1$ vector

5. Applications

Definition (Likelihood)

Under the normality assumption, the condional distribution of $\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\beta} + \varepsilon_t$ given $\mathbf{X}_t$ and the set of parameters $\boldsymbol{\beta}$ is normal

$$\mathbf{Y}_t | \mathbf{X}_t, \boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{X}_t \boldsymbol{\beta}, \boldsymbol{\Sigma}\right)$$

and the **likelihood** of the (realisations) sample $(\mathbf{Y}_1, .., \mathbf{Y}_T)$, denoted $\mathbf{Y}$, is:

$$L_T\left(\mathbf{Y} | \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \propto |\boldsymbol{\Sigma}|^{-T/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(\mathbf{Y}_t - \mathbf{X}_t \boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}_t - \mathbf{X}_t \boldsymbol{\beta}\right)\right)$$

By convention, we will denote $L_T\left(\mathbf{Y} | \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}\right) = L_T\left(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}\right)$

5. Applications

Definition (Joint posterior distribution)

Given a prior $\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, the **joint posterior distribution** of the parameters $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma} | \mathbf{Y}) = \frac{L_T(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\beta}, \boldsymbol{\Sigma})}{f(\mathbf{Y})}$$

or equivalently

$$\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma} | \mathbf{Y}) \propto L_T(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

Marginal posterior distribution

Given $\pi\left(\boldsymbol{\beta}, \boldsymbol{\Sigma} \middle| \mathbf{Y}\right)$, the marginal posterior distribution for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ can be obtained by integration

$$\pi\left(\boldsymbol{\beta} \middle| \mathbf{Y}\right) = \int \pi\left(\boldsymbol{\beta}, \boldsymbol{\Sigma} \middle| \mathbf{Y}\right) d\boldsymbol{\Sigma}$$

$$\pi\left(\boldsymbol{\Sigma} \middle| \mathbf{Y}\right) = \int \pi\left(\boldsymbol{\beta}, \boldsymbol{\Sigma} \middle| \mathbf{Y}\right) d\boldsymbol{\beta}$$

where $d\boldsymbol{\Sigma}$ and $d\boldsymbol{\beta}$ correspond to integrand defined for all the elements of $\boldsymbol{\Sigma}$ or $\boldsymbol{\beta}$ respectively

Bayesian estimates

Location and dispersion of $\pi\left(\boldsymbol{\beta}|\,\mathbf{Y}\right)$ and $\pi\left(\boldsymbol{\Sigma}|\,\mathbf{Y}\right)$ can be easily analysed to yield point estimates (quadratic loss) of the parameters of interest and measure of precision, comparable to those obtained by using classical approach to estimation. Especially:

$$\mathbb{E}\left(\pi\left(\boldsymbol{\beta}|\,\mathbf{Y}\right)\right) \qquad \mathbb{V}\left(\pi\left(\boldsymbol{\beta}|\,\mathbf{Y}\right)\right) \qquad \mathbb{E}\left(\pi\left(\boldsymbol{\Sigma}|\,\mathbf{Y}\right)\right)$$

5. Applications

Two problems arise:

1. The **numerical integration** of the marginal posterior distribution
2. The choice of the **prior**

5. Applications

Fact (Numerical integration)

In most of case, the **numerical integration** *of* $\pi\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right|\mathbf{Y})$ *may be difficult or even impossible to implement. For instance, if* $n=1$, $p=4$, *we have*

$$\pi\left(\boldsymbol{\Sigma}\right|\mathbf{Y}) = \int \pi\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right|\mathbf{Y})\, d\boldsymbol{\beta} = \int\int\int\int \pi\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right|\mathbf{Y})\, d\beta_1 d\beta_2 d\beta_3 d\beta_4$$

This problem, however, can often be solved by using numerical integration based on Monte Carlo simulations.

5. Applications

One particular **MCMC** (Markov-Chain Monte Carlo) estimation method is the **Gibbs sampler**, which a particular version of the **Metropolis-Hasting** algorithm (see section 5).

5. Applications

Definition (Gibbs sampler)

The **Gibbs sampler** is a recursive Monte Carlo method that allows to generate simulated values of $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ from the joint posterior distribution. This method requires only knowledge of the full **conditional** posterior distribution of the parameters of interest $\pi\left(\boldsymbol{\beta} \middle| \mathbf{Y}, \boldsymbol{\Sigma}\right)$ and $\pi\left(\boldsymbol{\Sigma} \middle| \mathbf{Y}, \boldsymbol{\beta}\right)$.

Definition (Gibbs algorithm)

Suppose that β and Σ are sacalr. The **Gibbs sampler** algorithm starts from arbitrary values $\left(\beta^0,\Sigma^0\right)$, and samples alternatively from the density of each element of the parameter vector, conditional of the value of the other element sampled in the previous iteration. Thus, the Gibbs sampler samples recursively as follows:

$$\beta^1 \text{ from } \pi\left(\beta|\, Y, \Sigma^0\right)$$

$$\Sigma^1 \text{ from } \pi\left(\Sigma|\, Y, \beta^1\right)$$

$$\beta^2 \text{ from } \pi\left(\beta|\, Y, \Sigma^1\right)$$

$$...$$

$$\beta^m \text{ from } \pi\left(\beta|\, Y, \Sigma^{m-1}\right)$$

5. Applications

Fact (Gibbs sampler)

The vectors (β^m, Σ^m) *form a Markov-Chain and for a sufficiently large number of iterations,* $m \geq M$*, can be regarded as draws from the* **true joint posterior distribution** $\pi(\beta, \Sigma | Y)$. *Given a large sample of draws from this limiting distribution,* $(\beta^m, \Sigma^m)_{m=M+1}^{G-M}$, *any* **posterior moment** *of* **marginal density** *of interest can then be easily estimated consistently with its corresponding* **sample average**. *For instance:*

$$\frac{1}{G} \sum_{m=M+1}^{G} \beta^m \rightarrow \mathbb{E}(\beta | Y)$$

Remark

- The process must be started somewhere, though it does not matter much where.
- Nonetheless, a **burn-in** period is required to eliminate the influence of the starting values. So, the first M values $\left(\beta^m, \Sigma^m\right)$ are discarded.

Example (Gibbs sampler)

We consider the bivariate normal distribution first. Suppose we wished to draw a random sample from the population

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Question: write a Matlab code to generate a sample of $n = 1,000$ observations of $(x_1, x_2)^\top$ with a Gibbs sampler.

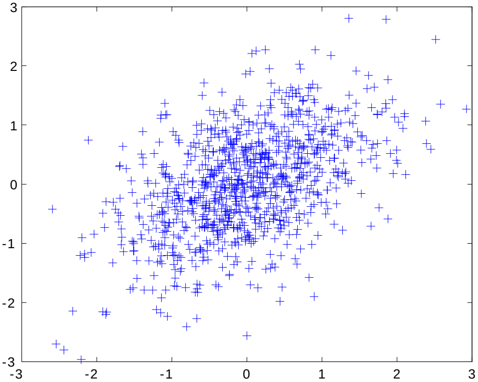
Solution

The Gibbs sampler takes advantage of the result

$$X_1 | \, x_2 \sim \mathcal{N}\left(\rho x_2, \left(1-\rho^2\right)\right)$$

$$X_2 | \, x_1 \sim \mathcal{N}\left(\rho x_1, \left(1-\rho^2\right)\right)$$

5. Applications



5. Applications

```
%=====================================================
% PURPOSE: Reproduce the Figure 18 of the Chapter 7
% Lecture: "Advanced Econometrics"
%-------------------------
% Author: Christophe Hurlin, University of Orleans
% Version: v1. June 2014
%=====================================================

clear all ; clc ; close all
rho=0.5;                % Correlation
burn=1000;              % Burn in period
n=1000;                 % Sample period
t=n+burn;               % total sample size
x1=NaN(t,1);            % Initialisation
x2=NaN(t,1);            % Initialisation
x1(1)=0;                % Initial value
x2(1)=0;                % Initial value
for i=2:t
   x1(i)=normrnd(rho*x2(i-1),1-rho^2);
   x2(i)=normrnd(rho*x1(i),1-rho^2);
end
x1=x1(burn+1:end);
x2=x2(burn+1:end);
plot(x1,x2,'+')
```

Priors

A second problem in implementing Bayesian estimation of VAR models is the choice of the **prior distribution** for the model's parameters

$$\pi\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)$$

5. Applications

We distinguish two types of priors:

1. Some priors lead to **analytical formula** for the posterior distribution
    1. Diffuse prior
    2. Natural conjugate prior
    3. Minnesoto prior
2. For other priors there is no analytical formula for the priors and the **Gibbs sampler** is required.
    1. Independent Normal-Wishart Prior

5. Applications

Definition (Full Bayesian estimation)

A **full Bayesian estimation** requires specifying prior distribution for the hyperparameters of the prior distribution of the parameters of interest.

5. Applications

Definition (Empirical Bayesian estimation)

The **empirical Bayesian estimation** is based on a preliminary estimation of the hyperparameters γ. These estimates could be obtained by OLS, maximum likelihood, GMM or other estimation method. Then, $\pi(\theta;\gamma)$ is substituted by $\pi(\theta;\widehat{\gamma})$. Note that the uncertainty in the estimation of the hyperparameters is not taken into account in the posterior distribution.

Empirical Bayesian estimation

In this section, we consider only **empirical Bayesian estimation** methods:

$$\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t$$

$$\boldsymbol{\varepsilon}_t \sim IIN\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$$

Notations:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{T} \mathbf{X}_t^{\top} \mathbf{X}_t \right)^{-1} \left(\sum_{t=1}^{T} \mathbf{X}_t^{\top} \mathbf{Y}_t \right) \quad \text{OLS estimator of } \boldsymbol{\beta}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T-k} \sum_{t=1}^{T} \left(\mathbf{Y}_t - \mathbf{X}_t \boldsymbol{\beta} \right)^{\top} \left(\mathbf{Y}_t - \mathbf{X}_t \boldsymbol{\beta} \right) \quad \text{OLS estimator of } \boldsymbol{\Sigma}$$

5. Applications

Matlab codes for Baysesian estimation of VAR models are proposed by Koop and Korobilis (2010)

1. **Analytical results**:
   1. Code **BVAR_ANALYT.m** gives posterior means and variances of parameters & predictives, using the analytical formulas.
   2. Code **BVAR_FULL.m** estimates the BVAR model combining all the priors discussed below, and provides predictions and impulse responses
2. **Gibbs sampler**: Code **BVAR_GIBBS.m** estimates this model, but also allows the prior mean and covariance.

Koop G. and Korobilis D. (2010), Bayesian Multivariate Time Series Methods for Empirical Macroeconomics

Sub-Section 4.1

Minnesota Prior

5. Applications

Fact (Stylised facts)

Literman (1986) specifies his prior by appealing to three statistical regularities of macroeconomic time series data:

(1) the trending behavior of macroeconomic time series

(2) more recent values of a series usually contain more information on the current value of the series than past values

(3) past values of a given variable contain more information on its current state than past values of other variables.

Litterman R. (1986), Forecasting with Bayesian Vector AutoRegressions, JBES, 4, 25-38.

5. Applications

A Bayesian researcher specify these regularities by assigning a probability distribution to the parameters in such way that:

1. the **mean** of the coefficients assigned to all lags other than the first one is equal to zero.
2. the **variance** of the coefficients depends inversely on the number of lags
3. the coefficient of variable j in equation g are assigned a lower prior **variance** than those of variable g.

These requirements will be controlled for by the **hyperparameters**

$$\pi = (\pi_1, .., \pi_d)^\top \Longrightarrow \text{prior on } \mathbf{B}_1, .., \mathbf{B}_p, \mathbf{D}, \mathbf{\Sigma}$$

Definition (Minnesota prior, Litterman 1986)

In the Minnesota prior, the variance covariance matrix $\mathbf{\Sigma}$ is assumed to be fixed and equal to $\widehat{\mathbf{\Sigma}}$. Denote $\boldsymbol{\beta}_k$ for $k = 1, .., n$, the vector of parameters of the k^{ith} equation, the corresponding prior distribution is:

$$\boldsymbol{\beta}_k \sim \mathcal{N}\left(\overline{\boldsymbol{\beta}}_k, \overline{\Omega}_k\right)$$

Remark

1. Given these assumptions, there is prior and posterior independence between equations.
2. The equations can be estimated separately.
3. By assuming $\overline{\Omega}_k = 0$, the posterior mean of $\boldsymbol{\beta}_k$ corresponds to the OLS estimator of $\boldsymbol{\beta}_k$.

5. Applications

Litterman (1986) then assigns numerical values to the hyperparameters given the previous stylised facts.

For the **prior mean:**

1. The traditional choice is to set $\overline{\beta}_k = 0$ for all the parameters if the k^{th} variable is a growth rate (random walk behavior).

2. If the variable is in level all the hyperparameters are set to 0, except the parameter associated to the first own lag.

5. Applications

The **prior covariance** matrix $\overline{\Omega}_k$ is assumed to be diagonal with elements $\omega_{i,i}$ for $i = 1, .., k$.

$$\omega_{i,i} = \begin{cases} \frac{a_1}{r^2} & \text{for coefficients on own lag } r = 1, .., p \\ \frac{a_2}{r^2} \frac{\sigma_{ii}}{\sigma_{jj}} & \text{for coefficients on lag } r \text{ of variable } j \neq i,\ r = 1, .., p \\ a_3 \sigma_{ij} & \text{for coefficients on exogenous variables} \end{cases}$$

This prior simplifies the complicated choice of fully specifying all the ele-ments of $\overline{\Omega}_k$ to choosing three scalars, a_1, a_2 and a_3.

5. Applications

Theorem (Posterior distribution)

If we assume Minnesota priors, the posterior distribution of $\boldsymbol{\beta}_k$ is:

$$\boldsymbol{\beta}_k | \mathbf{Y} \sim \mathcal{N}\left(\widetilde{\boldsymbol{\beta}}_k, \widetilde{\boldsymbol{\Omega}}_k\right)$$

$$\widetilde{\boldsymbol{\beta}}_k = \widetilde{\boldsymbol{\Omega}}_k \left(\widetilde{\boldsymbol{\Omega}}_k^{-1} \overline{\boldsymbol{\beta}}_k + \sigma_{k,k}^{-2} \mathbf{X}^\top \mathbf{Y}_g\right)$$

$$\widetilde{\boldsymbol{\Omega}}_k = \left(\overline{\boldsymbol{\Omega}}_k^{-1} + \sigma_{k,k}^{-2} \mathbf{X}^\top \mathbf{X}\right)^{-1}$$

Sub-Section 4.2

Diffuse Prior

5. Applications

Definition (Diffuse prior)

The Diffuse prior is defined as to be:

$$\pi\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-(n+1)/2}$$

5. Applications

Remark

$$\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(n+1)/2}$$

Contrary to Minnesota prior:

1. $\boldsymbol{\Sigma}$ is not assumed to be fixed
2. The equations are not independent

5. Applications

Theorem (Posterior distribution)

For a Diffusion prior distribution, the joint posterior distribution is given by:

$$\pi\left(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \mathbf{Y}\right) = \pi\left(\boldsymbol{\beta} \mid \mathbf{Y}, \boldsymbol{\Sigma}\right) \ \pi\left(\boldsymbol{\Sigma} \mid \mathbf{Y}\right)$$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\right)$$

$$\boldsymbol{\Sigma} \mid \mathbf{Y} \sim IW\left(\widehat{\boldsymbol{\Sigma}}, T - k\right)$$

where IW denotes the Inverted Wishart distribution.

Sub-Section 4.3

Natural conjugate prior

Definition (Natural conjugate prior)

The **natural conjugate prior** has the form:

$$\boldsymbol{\beta} | \boldsymbol{\Sigma} \sim \mathcal{N}\left(\overline{\boldsymbol{\beta}}, \boldsymbol{\Sigma} \otimes \overline{\boldsymbol{\Omega}}\right)$$

$$\boldsymbol{\Sigma} \sim IW\left(\overline{\boldsymbol{\Sigma}}, \alpha\right)$$

with $\alpha > n$ and IW denotes the Inverted Wishart distribution with a degree of freedom equal to s. The hyperparameters are $\overline{\boldsymbol{\beta}}$, $\overline{\boldsymbol{\Omega}}$, $\overline{\boldsymbol{\Sigma}}$ and α.

Theorem (Posterior distribution)

The posterior distribution associated to the natural conjugate prior is:

$$\beta | \boldsymbol{\Sigma}, \mathbf{Y} \sim \mathcal{N}\left(\widetilde{\beta}, \boldsymbol{\Sigma} \otimes \widetilde{\Omega}\right)$$

$$\boldsymbol{\Sigma} | \mathbf{Y} \sim IW\left(\widetilde{\boldsymbol{\Sigma}}, T + \alpha\right)$$

where

$$\widetilde{\beta} = \widetilde{\Omega}\left(\widetilde{\Omega}^{-1}\overline{\beta} + \mathbf{X}^{\top}\mathbf{X}\widehat{\beta}_{ols}\right)$$

$$\widetilde{\Omega} = \left(\overline{\Omega}^{-1} + \mathbf{X}^{\top}\mathbf{X}\right)^{-1}$$

Sub-Section 4.4

Independent Normal-Wishart Prior

5. Applications

Definition (Independent Normal-Wishart Prior)

The **independent Normal-Wishart prior** is defined as:

$$\pi\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}^{-1}\right) = \pi\left(\boldsymbol{\beta}\right)\ \pi\left(\boldsymbol{\Sigma}^{-1}\right)$$

where

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\overline{\boldsymbol{\beta}}, \overline{\boldsymbol{\Omega}}\right)$$

$$\boldsymbol{\Sigma}^{-1} \sim W\left(\overline{\boldsymbol{\Sigma}}^{-1}\right), \alpha$$

where $W\left(., \alpha\right)$ denotes the Wishart distribution with α degrees of freedom.

Remark

1. Note that this prior allows for the prior covariance matrix, $\overline{\mathbf{\Omega}}$, to be anything the researcher chooses, rather than the restrictive $\mathbf{\Sigma} \otimes \overline{\mathbf{\Omega}}$ form of the natural conjugate prior. For instance, the researcher could choose a prior similar in spirit to the Minnesota prior.

2. A noninformative prior can be obtained by setting

$$\overline{\boldsymbol{\beta}} = \overline{\mathbf{\Omega}} = \overline{\mathbf{\Sigma}}^{-1} = 0$$

5. Applications

Theorem (Posterior distribution)

In the case of independent Normal-Wishart prior, the joint posterior distribution can not be derived analytically. We can only derive the **conditional posterior distribution** *(used in the* **Gibbs sampler***)*

$$\beta|\,\Sigma, \mathbf{Y} \sim \mathcal{N}\left(\widetilde{\beta}, \Sigma \otimes \widetilde{\Omega}\right)$$

$$\Sigma|\,\mathbf{Y}, \beta \sim IW\left(\widetilde{\Sigma}, T + \alpha\right)$$

where

$$\widetilde{\beta} = \widetilde{\Omega}\left(\overline{\Omega}^{-1}\overline{\beta} + \sum_{t=1}^{T} \mathbf{X}_t^{\top} \Sigma^{-1} \mathbf{Y}_t\right)$$

$$\widetilde{\Omega} = \left(\overline{\Omega}^{-1} + \sum_{t=1}^{T} \mathbf{X}_t^{\top} \Sigma^{-1} \mathbf{X}_t\right)^{-1}$$

Sub-Section 5

Simulation methods

5. Simulation methods

Objectives

The objective of this section are the following:

1. Introduce the **Probability Integral Transform (PIT)** method
2. Introduce the **Accept-Reject (AR)** method
3. Introduce the **Importance sampling** method
4. Introduce the **Gibbs** algorithm
5. Introduce the **Metropolis Hasting** algorithm

5. Simulation methods

In bayesian econometrics, we may distinguish two case give the prior:

1. When we use a **conjugate prior**, the posterior distribution is sometimes "standard" (normal, gamma, beta etc..) and standard packages of the statistic softwares may be used to compute the density $\pi(\theta|y)$, its moments (for instance $\mathbb{E}\pi(\theta|y)$), the hpd etc.

2. In most of cases, the posterior density $\pi(\theta|y)$ is obtained numerically through **simulation methods.**

Sub-Section 5.1

Probability Integral Transform (PIT) method

Probability Integral Transform (PIT) method

- Suppose we wish to **draw** a sample of values from a continuous random variable X that has cdf $F_X(.)$, assumed to be nondecreasing.
- The PIT methods allows to generate a sample of X from a sample of random values issued from U where U has a **uniform distribution** over $[0,1]$:

$$U \sim U_{[0,1]}$$

Probability Integral Transform (PIT) method

Indeed, if we assume that the random variable U is a function of the random variable X such that:

$$U = F_X(X)$$

What is the distribution of U?

$$F_X(x) = \Pr(X \leq x) = \Pr(F_X(X) \leq F_X(x)) = \Pr(U \leq F_X(x))$$

Probability Integral Transform (PIT) method

$$F_X(x) = \Pr(X \leq x) = \Pr(F_X(X) \leq F_X(x)) = \Pr(U \leq F_X(x))$$

So, we have:

$$\Pr(U \leq F_X(x)) = F_X(x)$$

We know that if U has a uniform distribution then

$$\Pr(U \leq u) = u$$

So, U has a uniform distribution.

Definition (Probability Integral Transform (PIT))

If X is a continuous random variable with a cdf $F_X(.)$, then the **transformed (probability integral transformation)** variable $U = F_X(X)$ has a uniform distribution over $[0, 1]$.

$$U = F_X(X) \sim U_{[0,1]}$$

How to get a trial for X from a trial from U?

Definition (PIT algorithm)

In order to get a **realisation** x of the random variable of X with a cdf $F_X(.)$, from a realisation u of the variable U with $U \sim U_{[0,1]}$, the following procedure has to be adopted:

(1) Draw u from $U[0,1]$.

(2) Compute $x = F_X^{-1}(u)$

Example (PIT method)

Suppose we wish to draw a sample from a random variable with density function

$$f_X(x) = \begin{cases} \frac{3}{8}y^2 & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Question: Write a Matlab code a generate a sample $(x_1, .., x_n)$ with $n = 100$.

Solution (cont'd)

First, determine the cdf of X :

$$F_X(x) = \frac{3}{8}\int_0^x t^2 dt = \frac{1}{8}x^3$$

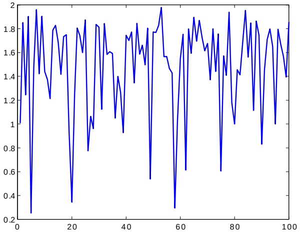
So, we have:

$$U = F_X(X) = \frac{1}{8}X^3 \sim U_{[0,1]}$$

Then, determine the probability inverse transformation:

$$X = F_X^{-1}(U) = 2U^{1/3}$$

Solution (cont'd)




```
%==================================================
% PURPOSE: Reproduce the Figure 12 of the Chapter 7
% Lecture: "Advanced Econometrics"
%-------------------------
% Author: Christophe Hurlin, University of Orleans
% Version: v1. June 2014
%==================================================
clear all ; clc ; close all

n=100;                    % Sample size
u=unifrnd(0,1,n,1);       % Sample of U(0,1)
x=2*u.^(1/3);             % Sample of x

plot(x,'Linewidth',1.5)
```

Remark

Note that a **multivariate** random variable cannot be simulated by this method, because its cdf is not one-to-one and therefore not invertible.

Remark

- An important application of this method is to the problem of sampling from a **truncated distribution**
- Suppose that X has a cdf $F_X(x)$ and that we want to generate restricted values of X such that

$$c_1 \leq X \leq c_2$$

Remark (cont'd)

The cdf of the truncated variable is

$$\frac{F_X(x) - F(c_1)}{F_X(c_1) - F_X(c_2)} \quad \text{for } c_1 \leq x \leq c_2$$

Then, we have:

$$U = \frac{F_X(X) - F(c_1)}{F_X(c_1) - F_X(c_2)} \sim U_{[0,1]}$$

and the truncated variable can be defined as:

$$X_{trunc} = F_X^{-1}\left(F_X(c_1) + U\left(F_X(c_2) - F_X(c_1)\right)\right)$$

Recommendation

1. Why using the **PIT method**?
   - In order to generate a sample of simulated values from a given distribution
   - This distribution is not available in my statistical software..
2. What are the prerequisites of the PIT?
   - The functional form of the cdf is known and we need to know $F^{-1}(x)$ ....

Sub-Section 5.2

Accept-reject method

Definition (Accept-reject method)

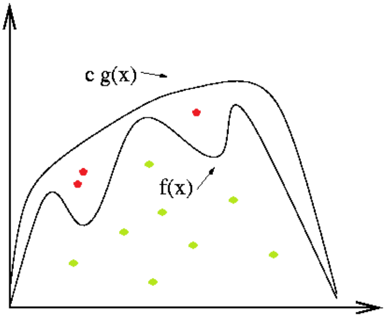
The **accept–reject (AR)** algorithm can be used to **simulate** values from a density function $f_X(.)$ if it is possible to simulate values from a density $g(.)$ and if a number c can be found such that

$$f_X(x) \leq c \times g(x)$$

for all x in the support of $f_X(.)$.

Definition (Target and source densities)

The density $f_X(.)$ is called the **target** density and $g(.)$ is called the **source** density.



c g(x)
f(x)

Posterior distribution

In the context of the Bayesian econometrics, we have:

$$\pi\left(\theta \mid y\right) \leq c \times g\left(\theta \mid y\right) \qquad \forall \theta \in \Theta$$

$$c = \sup_{\theta \in \Theta} \left\{ \frac{\pi\left(\theta \mid y\right)}{g\left(\theta \mid y\right)} \right\}$$

where:

1. **Target density** = posterior distribution $\pi\left(\theta \mid y\right)$
2. **Source density** = $g\left(\theta \mid y\right)$

Definition (AR algorithm, posterior distribution)

The Accept-Reject algorithm for the posterior distribution is the following:

(1) Generate a value m θ^s from $g\left(\theta|\,y\right)$.

(2) Draw a value u from $U_{[0,1]}$.

(3) Return θ^s as a draw from $\pi\left(\theta|\,y\right)$ if

$$u \leq \frac{\pi\left(\theta^s|\,y\right)}{cg\left(\theta^s|\,y\right)}$$

If not, reject it and return to step 1.(The effect of this step is to accept θ^s with probability $\pi\left(\theta^s|\,y\right)/cg\left(\theta^s|\,y\right)$.)

Example (AR algorithm)

We aim at simulating some realisations from a $\mathcal{N}(0,1)$ distribution (**target** distribution) with a **source** density given by a Laplace distribution with a pdf:

$$g(x) = \frac{1}{2} \exp(-|x|)$$

Question: write a Matlab code to simulate a sample of $n = 1,000$ realisations of the normal distribution.

Solution

In order to simplify the problem, we generate only positive values for x.

The source density can transformed as follows (exponential density with $\lambda = 1$).

$$g(x) = \exp(-x)$$

Since the normal and the Laplace are symmetric about zero, if the proposal $(x > 0)$ is accepted, it is assigned a positive value with probability one half and a negative value with probability one half.

Solution (cont'd)

The pdf of the target and source distributions are for $x > 0$:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad g(x) = \exp(-x)$$

Determination of c : determine the maximum value of $f(x) / g(x)$:

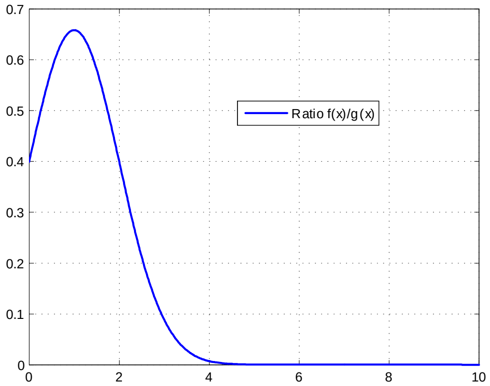
$$\frac{f(x)}{g(x)} = \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-\frac{x^2}{2}\right)}{\exp(-x)}$$

The maximum is reached for $x = 1$, then we have:

$$c = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}\right)$$

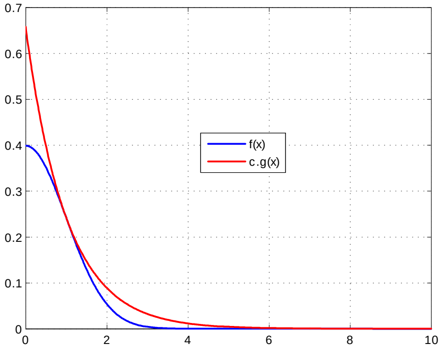
5. Simulation methods

5.2. Accept-reject method



5. Simulation methods

5.2. Accept-reject method



Solution (cont'd)

The **AR algorithm** is the following:

1. Generate x from an exponential distribution with parameter $\lambda = 1$.
2. Generate u from an uniform distribution $U_{[0,1]}$
3. If

$$u \leq \frac{\phi\left(x\right)}{c \times \exp\left(-x\right)} = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(1/2\right) \exp\left(-x\right)} = \exp\left(x - \frac{x^2}{2} + x\right)$$

then return x if $u > 1/2$ and $-x$ if $u \leq 1/2$. Otherwise, reject x.

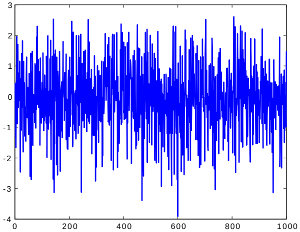
5. Simulation methods

5.2. Accept-reject method

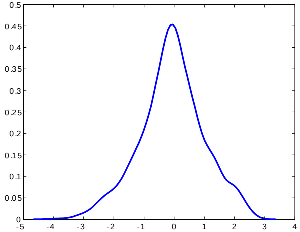
```
%=====================================================
% PURPOSE: Reproduce the Figure 14 of the Chapter 7
% Lecture: "Advanced Econometrics"
%-------------------------
% Author: Christophe Hurlin, University of Orleans
% Version: v1. June 2014
%=====================================================
clear all ; clc ; close all

% Simulation
n=100;                      % Sample size
c=exp(0.5)*sqrt(2*pi);      % constant c
comp=0;                     % compteur
X=NaN(n,1);                 % Sample
while comp<n
    u=unifrnd(0,1);         % Draw from a U(0,1)
    x=exprnd(1);            % Draw from a exp(1) distribution
    if u<exp(x-x^2/2+1/2)
        comp=comp+1;
        if u>0.5
            X(comp)=x;
        else
             X(comp)=-x;
        end
    end
end
```

Generated sample



Kernel density estimates



Fact (Unormalised posterior distribution)

An interesting feature of the AR algorithm is that is allows to simulate some values for θ (from the posterior distribution $\pi(\theta|\, y)$) by only using the **unormalised posterior**

$$\pi(\theta|\, y) = \frac{f(y|\,\theta)\,\pi(\theta|\, y)}{f(y)}$$

$$\pi(\theta|\, y) \propto f(y|\,\theta)\,\pi(\theta|\, y)$$

Intuition

$$\pi\left(\theta | y\right) = \underbrace{\frac{1}{f\left(y\right)}}_{\text{unknown}} \times \underbrace{f\left(y | \theta\right) \pi\left(\theta | y\right)}_{\text{known}}$$

Assume that $f\left(y | \theta\right) \pi\left(\theta | y\right)$ is known, but $1/f\left(y\right)$ is unknow.

If a value of θ generated from $g\left(\theta | y\right)$ is accepted with probability $f\left(y | \theta\right) \pi\left(\theta | y\right) / cg\left(\theta | y\right)$, the accepted values of θ are a sample from $\pi\left(\theta | y\right)$.

This method can therefore be used even if the **normalizing constant of the target distribution is unknow**

Example (AR and posterior distribution)

Consider an *i.i.d.* sample $(Y_1, .., Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$ with $\theta = 0.3$. We assume that the (uninformative) prior distribution for θ is an uniform distribution over $[0, 1]$. Then, we have

$$\pi(\theta | y) = \frac{L_n(\theta; y_1, .., y_n) \times \pi(\theta)}{f_{Y_1,..Y_n}(y_1, .., y_n)} = \frac{\theta^{\Sigma y_i}(1-\theta)^{\Sigma(1-y_i)}}{\int_0^1 \theta^{\Sigma y_i}(1-\theta)^{\Sigma(1-y_i)} d\theta}$$

We can use the pdf of the unormalised posterior $\widetilde{\pi}(\theta | y)$ to simulate some values $(\theta_1, ..\theta_S)$ from the posterior distribution

$$\widetilde{\pi}(\theta | y) = \theta^{\Sigma y_i}(1-\theta)^{\Sigma(1-y_i)}$$

Recommendation

1. Why using the **Accept-Reject method**?
   - In order to generate a sample of simulated values from a given distribution.
   - This distribution is not available in my statistical software.
   - To generate samples of θ from the **unormalised posterior distribution**

2. What are the prerequisites of the PIT?
   - The functional form of the **pdf** of the target distribution is known

Sub-Section 5.3

Importance sampling

Suppose that one is interested in calculating the value of the integral

$$I = \mathbb{E}\left(h\left(\theta\right) \middle| y\right) = \int_{\Theta} h\left(\theta\right) \pi\left(\theta \middle| y\right) d\theta$$

where $h\left(\theta\right)$ is a continuous function.

Example (Importance sampling)

Suppose that we want to compute the expectation and the variance of the posterior distribution

$$\mathbb{E}\left(\theta \mid y\right) = \int_{\Theta} \theta \; \pi\left(\theta \mid y\right) d\theta$$

$$\mathbb{E}\left(\theta^2 \mid y\right) = \int_{\Theta} \theta^2 \; \pi\left(\theta \mid y\right) d\theta$$

$$\mathbb{V}\left(\theta \mid y\right) = \mathbb{E}\left(\theta^2 \mid y\right) - \mathbb{E}^2\left(\theta \mid y\right)$$

$$I = \int_{\Theta} h(\theta)\, \pi(\theta|\, y)\, d\theta$$

Consider a **source density** $g(\theta|\, y)$ that is easy to sample from and which is a close match to $\pi(\theta|\, y)$. Write:

$$I = \int_{\Theta} \frac{h(\theta)\, \pi(\theta|\, y)}{g(\theta|\, y)} g(\theta|\, y)\, d\theta$$

This integral can be approximated by drawing a sample of G values from $g(\theta|\, y)$ denotes $\theta_1^g, .., \theta_G^g$ and computing

$$I \simeq \frac{1}{G} \sum_{i=1}^{G} h\left(\theta_i^g\right) \frac{\pi\left(\theta_i^g \,|\, y\right)}{g\left(\theta_i^g \,|\, y\right)}$$

Definition (Importance sampling)

The moment of the posterior distribution

$$\mathbb{E}\left(h\left(\theta\right)\middle|\, y\right) = \int_{\Theta} h\left(\theta\right)\pi\left(\theta\middle|\, y\right) d\theta$$

can be approximated by drawing G realisations $\theta_1^g, .., \theta_G^g$ from a source density $g\left(\theta_i^g \,\middle|\, y\right)$ and computing

$$\mathbb{E}\left(h\left(\theta\right)\middle|\, y\right) \simeq \frac{1}{G}\sum_{i=1}^{G} h\left(\theta_i^g\right)\frac{\pi\left(\theta_i^g \,\middle|\, y\right)}{g\left(\theta_i^g \,\middle|\, y\right)}$$

This expression can be regarded as a weighted average of the $h\left(\theta_i^g\right)$ where **importance weights** are $\pi\left(\theta_i^g \,\middle|\, y\right)/g\left(\theta_i^g \,\middle|\, y\right)$

5. Simulation methods

5.3. Importance sampling

Example (Truncated exponential)

Consider a continuous random variable X with an exponential distribution $X \sim \exp(1)$ **truncated** to $[0, 1]$. We want to approximate

$$\mathbb{E}\left(\frac{1}{1+X^2}\right)$$

by the **importance sampling** method with a source density $B(2, 3)$ because it is defined over $[0, 1]$ and because, for this choice of parameters, the match between the beta function and the target density is good.

Question: write a Matlab code to approximate this integral. Compare it to the value obtained by numerical integration.

Solution

For an exponential distribution with a rate parameter of 1 we have

$$f(x) = \exp(-x) \qquad F(x) = 1 - \exp(-x) \; forx > 0$$

If this density is truncated over $[0, 1]$ (truncated at right) we have

$$\pi(x) = \frac{f(x)}{F(1)} = \frac{\exp(-x)}{1 - \exp(-1)}$$

So, we aim at computing:

$$\mathbb{E}\left(\frac{1}{1 + X^2}\right) = \int_0^1 \frac{1}{1 + x^2} \frac{\exp(-x)}{1 - \exp(-1)} dx$$

Solution (cont'd)

The importance sampling algorithm is the following:

1. Generate a sample of G values $x_1, .., x_G$ from a Beta distribution $B(2,3)$
2. Compute

$$\mathbb{E}\left(\frac{1}{1+X^2}\right) \simeq \frac{1}{G}\sum_{i=1}^{G} \underbrace{\left(\frac{1}{1+x_i^2}\right)}_{h(x_i)} \underbrace{\left(\frac{\exp(-x_i)}{1-\exp(-1)}\right)}_{\pi(x_i)} \underbrace{\frac{1}{g_{2,3}(x_i)}}_{g(x_i)}$$

where $g_{\alpha,\beta}(x_i)$ is the pdf of the $B(\alpha,\beta)$ distribution evaluated at x

$$g_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

and $\mathrm{B}(\alpha,\beta)$ is the beta function.

```
%====================================================
% PURPOSE: Reproduce the Figure 15 of the Chapter 7
% Lecture: "Advanced Econometrics"
%-------------------------
% Author: Christophe Hurlin, University of Orleans
% Version: v1. June 2014
%====================================================

clear all ; clc ; close all

G=100000;
x=betarnd(2,3,G,1);                    % Trials from the s
I=mean(1./(1+x.^2).*exp(-x)/(1-exp(-1))./betapdf(x,2,3)
disp(' '),disp('Importance sampling')
disp(I)
Inum=quad('expectation',0,1);
disp(' '),disp('Numerical integration')
disp(Inum)

function f=expectation(x)
f=1./(1+x.^2).*exp(-x)/(1-exp(-1));
```

Results

```
Importance sampling
    0.8406

Numerical integration
    0.8302
```

Recommendation

1. Why using the **Importance sampling method**?
   - In order to compute the **moments** of a given distribution (typically the **posterior** distribution).
2. What are the prerequisites of the PIT?
   - The functional form of the **pdf** of the target distribution is known

End of Chapter 7

Christophe Hurlin (University of Orléans)