Bayesian Time-Series Econometrics

Book 1 - theory

Romain Legrand

First edition

Bayesian Time-Series Econometrics

Cover illustration: Thomas Bayes (d. 1761) in Terence O'Donnell, History of Life Insurance in Its Formative Years (Chicago: American Conservation Co:, 1936), p. 335.

To my wife, Mélanie.

To my sons, Tristan and Arnaud.

Contents

PART I

Bayesian statistics

CHAPTER 1

Bayesian and frequentist approachs

Jourdain: There is this person of great quality and I want you to help me to write a short love note which I can drop at her feet.
Philosophy Master: Fine. Do you wish to write to her in verse?
Jourdain: No, no poetry.
Philosophy Master: So you desire prose.
Jourdain: Oh, no! I don't want prose or poetry.
Philosophy Master: It must be one or the other.
Jourdain: Why?
Philosophy Master: Because there is no other way to express oneself but through prose or verse. Whatever is not prose, is poetry and whatever is not poetry is prose.
Jourdain: When I talk what's that then?
Philosophy Master: Prose.
Jourdain: When I say, Nicole! Bring me my slippers, is that prose.
Philosophy Master: Yes, sir.
Jourdain: So I have been speaking prose for years without even knowing it! What a Master you are.

Molière *, The Bourgeois Gentleman*

1.1 Introduction

It may sound surprising to start a book on Bayesian statistics with an extract from a play by Molière. Yet we can draw a parallel between this dialog and the statistical approach discussed in this book. In this extract, Mr. Jourdain (the main character of "The Bourgeois Gentleman") first discovers that speaking is formally known as "prose". Moreover, he learns that there exist in fact two ways to express oneself: prose, and poetry. The same goes for statistics. Most statisticians follow an approach formally known as the "frequentist" approach, without being aware of it. In addition, there exist in fact two different approaches to statistics: the frequentist approach, and the Bayesian approach[1].

This first chapter introduces the two approaches and highlights their main differences. It does not yet deal with the technicalities of the subject, left to the incoming chapters. Rather, it develops the fundamental concepts in a purposely informal way in order to set the terms of the debate.

1.2 Fundamental concepts

In general, any statistical exercise is concerned with the outcome of some random experiment.

definition 1.1: a **random experiment** is a process whose outcome is uncertain, and can be known only once it is realized and observed.

[1] There exist actually more than two approaches to statistics, such as the symmetric and logical approaches. Those alternative approaches are not of interest for this book and are not developed further: see for instance Poirier (1995) for more details.

Here are a few examples of random experiments.

example 1.1: the outcome of a coin flip.

example 1.2: the number of cars sold in a month at a car retailer's.

example 1.3: the market return of a stock at the New York Stock exchange.

To understand the behaviour of a random experiment, the statistician creates a model which replicates its statistical properties. This model typically depends on a number of parameters, denoted by θ.

definition 1.2: a **statistical model** is a model that describes the underlying process generating the data. It is indexed by a family of **parameters** θ which determine the behaviour of the model.

This can be illustrated with the simple examples introduced above:

example 1.1 (continued): to model the outcome of a coin flip, the statistician may use a Bernoulli distribution with probability of success p. In this case, p represents the parameter of the model, so that $\theta = \{p\}$.

example 1.2 (continued): to model the number of cars sold during a month at a car retailer's, the statistician may use a Poisson experiment with intensity λ, which represents the mean of the process. In this case, λ represents the parameter of the model, so that $\theta = \{\lambda\}$.

example 1.3 (continued): to model the market return of a stock, the statistician may use a normal distribution, where the mean μ represents the expected return of the stock and the variance σ represents its volatility. Here μ and σ represent the parameters of the model, and $\theta = \{\mu, \sigma\}$.

Because the parameters determine the behavior of the model, they represent the fundamental object of interest. They thus constitute the values that the statistician wants to estimate. In this respect, the main differences between the frequentist approach and the Bayesian approach arise in the way θ is considered, and as a consequence in the methodologies employed to estimate it.

1.3 The frequentist approach

When statisticians talk about "statistics", they usually mean the frequentist approach. Frequentist statisticians believe in random experiments which can be repeated. They assume that with a sufficiently large number of repetitions, probabilities can be deduced from observed frequencies, hence the name "frequentist". Concretely, for a given a random experiment repeated n times, and a possible outcome A of this random experiment observed m times over the n trials, the frequentist approach defines the probability of outcome A as:

definition 1.3: $P(A) = \underset{n \to \infty}{lim} \frac{m}{n}$

For a model parameters, definition 1.3 yields two main implications. First, for any statistical experiment and any outcome, there exists a unique and well defined probability which obtains as a limiting case of observed frequencies. Therefore, any parameter θ involved in a statistical model is also characterised by a unique and well-defined value. This value can be calculated exactly as long as one is capable of repeating the underlying random experiment an infinite number of times. Thus, in frequentist statistics, θ is treated as a fixed quantity.

The second implication of definition 1.3 is that under the frequentist approach probabilities are deduced from observed outcomes. The only source of information for a frequentist statistician thus consists in the data collected for the experiment. Because this data is assumed to be generated by the statistical model, any observation results from the parameter θ which determines the behaviour of the model. Following, the data constitutes the basis of any estimation process for θ.

In this respect, a fundamental object of interest is known as the likelihood function.

definition 1.4: let y denote the sample of data collected by the statistician, and let θ denote the model parameters; the **likelihood function** , denoted by $f(y|\theta)$, represents the density function of the observed data y, for a given value of θ.

The likelihood function indicates how likely the observed data is for a given value of θ. A high value for $f(y|\theta)$ indicates that it is plausible to obtain the observed data with the given θ. Conversely, a low value for $f(y|\theta)$ implies that the selected θ makes the observed data unlikely to occur. A natural step then consists in estimating θ by choosing the value which makes the observed data most likely. This is the principle underlying the maximum likelihood methodology.

definition 1.5: the **maximum likelihood** estimation methodology consists in finding the value $\hat{\theta}$ which maximises the likelihood function $f(y|\theta)$. $\hat{\theta}$ is then called a **point estimate** for θ.

Given the information contained in the data, the point estimate $\hat{\theta}$ represents the best guess one can produce about the true parameter value θ. When the data sample is not infinite, some uncertainty exists about the parameter value. One may then want to construct confidence intervals on the parameter value.

definition 1.6: a **confidence interval** is an interval of values that contain the true value θ with high probability, set by the statistician.

Alternatively, a hypothesis test can be conducted on the parameter value.

definition 1.7: a **hypothesis test** is an inference procedure establishing whether a default hypothesis about θ called the null hyposthesis is true. If there is sufficient evidence, the null hypothesis is rejected in favor of the so-called alternative hypothesis.

1.4 The Bayesian approach

The frequentist approach defines probabilities as a limiting case of experiments repeated an infinite number of times, as stated by definition 1.3. By contrast, the Bayesian approach considers that in practical situations random experiments cannot be repeated an infinite number of times, or cannot be repeated at all. For instance, the weather in Washington DC on July 4th 2000 is not a repeatable random experiment since July 4th 2000 only occured once. Certain experiments can be repeated, such as the number of customers visiting a local grocery store during a day. They may however not be repeated an infinite number of times. Even if the number of repetitions tends to infinity, the random experiment being repeated may not be exactly the same. A grocery store updates its prices and line of products from time to time. It also runs sales, recruits new staff, modifies it display, and so on. These differences affect the number of customers, and alter the underlying random experiment.

For these reasons, Bayesian statistics considers that any random experiment involves fundamental uncertainty, and that it is impossible to get rid of this uncertainty. Statisticians must then estimate probabilities not only from the information carried by the data, but also from personal probability assessments which reflect their subjective beliefs about the outcome of the experiment.

This has two main implications. First, the fundamental uncertainty implies that θ cannot be considered as a fixed quantity anymore. Instead, θ must be treated as a random variable, and assigned a probability distribution. As a consequence, the object of interest for the statistician is not anymore the fixed value of θ (which is impossible to determine), but the probability distributions of the parameters θ.

Second, the fundamental uncertainty implies that the data resulting from observation does not constitute a sufficient source of information. Because there can only be a finite number of data observations, and because these observations are generated by different realisations of θ from its probability distribution, it is impossible to eliminate the uncertainty associated with θ. As a consequence, the data can only represent part of the information involved in the estimation process. It must be supplemented with additional information provided by the statistician, which represents his personal belief about the random experiment.

Concretely, it implies that the likelihood function which represents the information contained in the data is not sufficient anymore to obtain an estimate of θ. The estimation process must also involve the personal assessment of the statistician about the distribution function of θ, which is known as the prior distribution.

definition 1.8: the **prior distribution**, denoted by $\pi(\theta)$, is the distribution function which represents the personal belief of the statistician about the distribution of the parameters of interest θ.

Because the data is not the only source of information under the Bayesian approach, maximum likelihood does not constitute a suitable methodology of estimation. To account for both the data information contained in the likelihood function $f(y|\theta)$ and the personal information contained in the prior distibution $\pi(\theta)$, a Bayesian statistician will apply a methodology known as Bayes Rule. This methodology produces what is known as the posterior distribution for θ, which is a full distribution function reflecting both the information contained in the data and the subjective information introduced by the statistician.

definition 1.9: the **posterior distribution**, denoted by $\pi(\theta|y)$, is the distribution function of the parameter of interest θ obtained by the application of Bayes rule. It is obtained by combining the likelihood function $f(y|\theta)$ and the prior distribution $\pi(\theta)$, and represents the distribution of θ conditional on having observed the data y.

Unlike the frequentist approach for which the estimation produces a single point estimate $\hat{\theta}$, the Bayesian approach results in a full posterior distribution $\pi(\theta|y)$. This posterior distribution summarizes all the relevant information about θ and represents the workhorse of Bayesian statistics. It can be used for instance to generate credibility intervals.

definition 1.10: a **credibility interval** is an interval over a posterior distribution within which a parameter value falls with a certain probability.

The credibility interval represents the counterpart of the frequentist confidence interval, but its philosophy is different. A confidence interval treats the parameter as fixed, creating an interval that hopefully contains the true value. A credibility interval treats the parameter as random, and defines an interval that contains its values with some given probability.

It is also possible to conduct hypothesis tests in a Bayesian framework.

definition 1.11: a Bayesian hypothesis test consists in a comparison of the posterior probabilities under the null and alternative hypotheses. This comparison is summarized by a single value known as the **Bayes factor**.

Unlike the frequentist approach which aims at testing for the true parameter value, a Bayesian hypothesis test determines which model is more likely under the null and alternative hypotheses about θ.

1.5 Summary

This chapter has underlined the fundamental differences between the frequentist and Bayesian approaches. Those differences are summarised in Table 5.1 for convenience.

	frequentist	Bayesian
random experiments	can be infinitely repeated	cannot be infinitely repeated
certainty	certainty with infinite repetitions	fundamental uncertainty
parameter θ	unique, fixed value	random variable
object of interest	true value of θ	probability distribution of θ
relevant information	observed data only	observed data and personal information
source of information	likelihood function $f(y\|\theta)$	likelihood function $f(y\|\theta)$ and prior distribution $\pi(\theta)$
estimation technique	maximum likelihood	Bayes rule
estimate for θ	point estimate	posterior distribution
intervals	confidence interval	credibility interval
hypothesis test	decide of true value	decide of best model

Table 1.1: Main differences between the frequentist and Bayesian approaches

The incoming chapters initiate the technical part of the discussion. Chapter 2 introduce basic probability concepts and derives Bayes rule in the context of events and random variables. Chapter 3 then provides some practice on the subject through a set of simple examples. Chapter 4 discusses some important additional aspects of Bayesian priors and posteriors. Chapter 5 concludes the first part by providing further insight on the properties of Bayesian estimates.

CHAPTER 2

Bayes rule

This chapter introduces Bayes rule, a result that constitutes the foundation of the whole field of Bayesian statistics. It does so first in the simple context of events, then extends to the more general notion of random variables. The presentation remains informal, only dealing with the aspects required to understand the incoming chapters. For this reason, the technicalities associated with formal probabilistic theory are left aside.

2.1 Probabilities

Probabilities are fundamentally concerned with random experiments and their outcomes. The first element of interest is thus the set of possible outcomes, known as the sample space.

> **definition 2.1:** the **sample space**, denoted by Ω, is the set of all possible outcomes of a random experiment. A subset of the sample space is called an **event**.

To illustrate this definition, let's take a look at some simple examples:

example 2.1: consider the random experiment "roll a 6-face die". Then the sample space is:
$\Omega = \{1,2,3,4,5,6\}$.
The subsets $A = \{2,4,6\}$, $B = \{4,5,6\}$ and $C = \{1\}$ are examples of events. They respectively correspond to: "the outcome of the roll is an even number", "the outcome of the roll is a number greater than 3", and "the outcome of the roll is 1".

example 2.2: consider the random experiment "pick a random number between 0 and 1". Then the sample space for this experiment is the closed interval $\Omega = [0,1]$.
The subsets $A = [0.1, 0.3]$ and $B = [0.5, 0.5]$ are examples event. They correspond to: "the picked number is comprised between 0.1 and 0.3" and "the picked number is 0.5".

Once equiped with a sample space, we associate probabilities to the events of interest by the way of a function known as a probability measure.

> **definition 2.2:** a **probability measure** is a function $\mathbb{P}(A)$ which associates a probability to each event A.

For instance:

example 2.1 (continued): if the die is balanced, each face has a $1/6$ probability to show up. So for an event A containing $|A|$ outcomes ($|A|$ denotes the cardinality, or number of elements of A), we want the probability to be $\mathbb{P}(A) = |A|/6$.
So for instance, considering $A = \{2,4,6\}$, we obtain $\mathbb{P}(A) = |A|/6 = 3/6 = 1/2$. Thus the probability of obtaining an even number from the roll is 1/2, as expected.

example 2.2 (continued): assume each number in $[0,1]$ is equally likely to be picked by the computer. This is a uniform setting in which the probability of any interval $[a,b]$ is simply equal to its length $(b-a)$. Then $\mathbb{P}(A) = b-a$.
So for instance, considering $A = [0.1, 0.3]$, we obtain $\mathbb{P}(A) = 0.3 - 0.1 = 0.2$. The probability of picking a number in the interval $[0.1, 0.3]$ is 0.2.

2.2 Bayes rule for events

To obtain Bayes rule, it is first necessary to introduce the concept of conditional probabilities.

definition 2.3: let A and B be two events on some sample space; the **conditional probability** of A given B, denoted by $\mathbb{P}(A|B)$ is given by:
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

The conditional probability $\mathbb{P}(A|B)$ must be understood as: "what is the probability that event A occurs, given that event B has occurred?". Figure 2.1 helps to make sense of the conditional probability formula in definition 2.3. If event B has occured, then clearly event A can only occur on the intersection portion $A \cap B$. However, we cannot use directly the probability $\mathbb{P}(A \cap B)$ since the sample space to consider is not the whole of Ω anymore, but is now restricted to event B. The conditional probability must thus be defined as the ratio of the grey area (the probability $\mathbb{P}(A \cap B)$) over the surface of event B (the probability $\mathbb{P}(B)$).

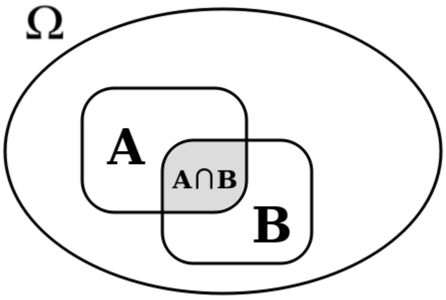


Figure 2.1: A representation of conditional probabilities with an Euler diagram

This can be illustrated with our usual examples.

example 2.1 (continued): consider the events $A = \{2,4,6\}$ (the outcome of the die roll is an even number) and $B = \{4,5,6\}$ (the outcome of the die roll is greater than 3). The conditional probability $\mathbb{P}(A|B)$ corresponds to "what is the probability that the outcome of the roll is even, given that it is greater than 3?"
We have $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 1/2$, $A \cap B = \{4,6\}$, $\mathbb{P}(A \cap B) = 1/3$, and $\mathbb{P}(A|B) = (1/3)/(1/2) = 2/3$
The unconditional probability of obtaining an even number $\mathbb{P}(A) = 1/2$ has been updated into the conditional probability $\mathbb{P}(A|B) = 2/3$ with additional information provided from observing B.

example 2.2 (continued): let $A = [0.1, 0.3]$ and $B = [0.2, 0.4]$
We have $\mathbb{P}(A) = 0.2$, $\mathbb{P}(B) = 0.2$, $A \cap B = [0.2, 0.3]$, $\mathbb{P}(A \cap B) = 0.1$, and $\mathbb{P}(A|B) = 0.1/0.2 = 1/2$
The unconditional probability of drawing a random number between 0.1 and 0.3 is $\mathbb{P}(A) = 1/5$, but increases to $\mathbb{P}(A|B) = 1/2$ if it is observed that the outcome is comprised between 0.2 and 0.4.

A first version of Bayes rule can now be obtained directly from the definition of conditional probability. Indeed, definition 2.3 implies $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ and $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$. Substituting the latter in the former yields Bayes rule:

definition 2.4: let A and B be two events on some sample space; **Bayes rule** is given by:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

The left-hand side of Bayes rule is the conditional probability $\mathbb{P}(A|B)$ which represents the probability of event A once B has been observed. It is equal to the right-hand side made of three components: the unconditional probability $\mathbb{P}(A)$, which represents the estimate of the probability of event A before event B is observed; the probability $\mathbb{P}(B)$, which corresponds to the additional information obtained from observing B; and the conditional probability $\mathbb{P}(B|A)$, which indicates how likely it is to observe B if event A occurs. Bayes rule then says that $\mathbb{P}(A|B)$ is equal to the unconditional probability $\mathbb{P}(A)$ updated by the additional evidence $\mathbb{P}(B|A)/\mathbb{P}(B)$.

example 2.1 (continued): let $A = \{2,4,6\}$ and $B = \{4,5,6\}$.
One has $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 1/2$, $\mathbb{P}(A \cap B) = 1/3$, and $\mathbb{P}(B|A) = 2/3$
Hence $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(B)/\mathbb{P}(A) = (2/3) \times (1/2)/(1/2) = 2/3$

example 2.2 (continued): let $A = [0.1, 0.3]$ and $B = [0.2, 0.4]$
We have $\mathbb{P}(A) = 0.2$, $\mathbb{P}(B) = 0.2$, $\mathbb{P}(A \cap B) = 0.1$, and $\mathbb{P}(B|A) = 1/2$
Hence $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B) = 1/2 \times 0.2/0.2 = 1/2$

2.3 Random variables

A preliminary version of Bayes rule has been introduced in the simple case of events. In practical applications however, Bayes rule is often used in the more general context of random variables.

definition 2.5: let Ω be some sample space; a **random variable** is a function $X(\omega)$ which associates a value to each outcome ω of the sample space.

Informally, a random variable can be seen as a function providing an interpretation to the outcome of a random experiment through the value it returns. For instance:

example 2.1 (continued): let X be the random variable defined as $X(\omega) = 1$ if $\omega = 2,4,6$, and $X(\omega) = 0$ otherwise. Its interpretation is: "observe whether the outcome of the roll was even".
$Z(\omega) = \omega$ is also a random variable. Its interpretation is simply: "reports the outcome of the die roll".

example 2.2 (continued): let $X(\omega) = 3\omega - 2$.
X is a random variable that can be interpreted as a lottery where the player pays 2 to play, then gains 3 times a random amount ω comprised between 0 and 1.

Random variables can be of two kinds. If it is possible to count the values a random variable can take, it is said to be discrete. If counting the values is impossible, typically because the random variables take values on some continuous interval, it is said to be continuous.

definition 2.6: a random variable X is called **discrete** if it takes only a finite or countable number of values. By contrast, a random variable X is said to be **continuous** if its values represent some continuous interval in $\mathbb{R}$.

The difference is best understood with the usual set of examples:

example 2.1 (continued): let X be the random variable defined as $X(\omega) = 1$ if $\omega = 2,4,6$, and $X(\omega) = 0$ otherwise. Then X is discrete since it takes a countable number of values (the two values 0 and 1).

example 2.2 (continued): let $X(\omega) = 3\omega - 2$. X takes values on the continuous interval $[-2,1]$ and is thus a continuous random variable.

So far our definition of random variables does not involve probabilities. The way probabilities are defined for random variables depends on their types. Because a discrete random variable can take only a countable number of values, it is possible to assign directly a probability to each value. This yields the concept of probability mass function.

definition 2.7: let X be a discrete random variable; then X has a **probability mass function** $f(x)$ such that $f(x) = \mathbb{P}(X = x)$, with $\sum_x f(x) = 1$.

The first statement defines the probability associated to each x value, while the second statement is just the classical condition that probabilities over all possible values should sum up to 1.

example 2.1 (continued): let X be the random variable defined as $X(\omega) = 1$ if $\omega = 2,4,6$, and $X(\omega) = 0$ otherwise. Its probability mass function is given by $f(1) = \mathbb{P}(X = 1) = \mathbb{P}(\{2,4,6\}) = 1/2$, and $f(0) = \mathbb{P}(X = 0) = \mathbb{P}(\{1,3,5\}) = 1/2$. Also, $f(1) + f(0) = 1/2 + 1/2 = 1$.

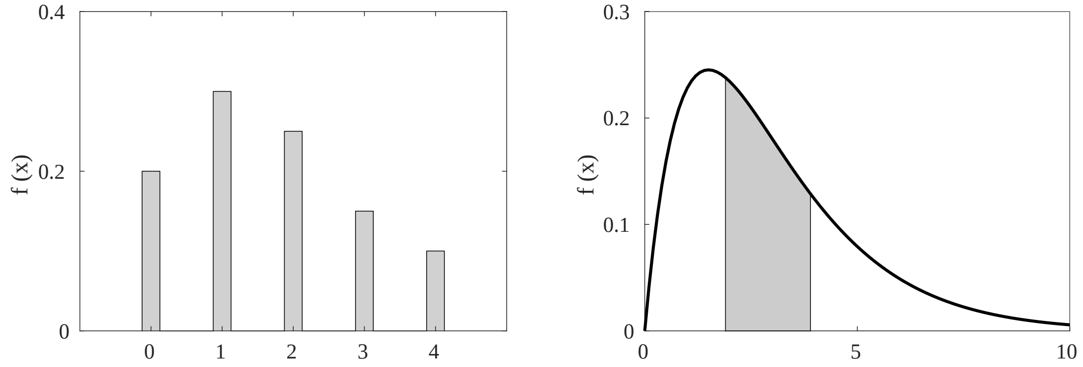
By contrast, continuous random take an uncountable number of possible values so that the probability of obtaining any single value is 0. Probabilities then only make sense over continuous intervals, which yields the notion of probability density function.

definition 2.8: let X be a continuous random variable; then X has a **probability density function** $f(x)$ such that: $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx$, with $\int_{-\infty}^{\infty} f(x)dx = 1$.

For instance:

example 2.2 (continued): let $X(\omega) = 3\omega - 2$. It can be shown that its probability density function is $f(x) = 1/3$, so that for instance $\mathbb{P}(0 \leq X \leq 1) = \int_0^1 1/3dx = 1/3$. Also, $\int_{-\infty}^{\infty} f(x)dx = \int_{-2}^{1} 1/3dx = 1$.

The conceptual difference between probability mass functions and probability density functions is illustrated by Figure 2.2. For the discrete random variable on the left panel, probabilities are attributed to each value of the random variable. For the continuous random variable on the right panel, probabilities are only defined by integrating over intervals (calculating the surface under the curve, such as the grey area).



(a) A probability mass function (b) A probability density function

Figure 2.2: Examples of mass and density functions

2.4 Bayes rule for random variables

The previous sections focused on individual random variables. In practice however, most statistical models involve more than one random variable at a time. We then want to consider probabilities not only for single random variables, but also for groups or random variables considered jointly. For instance, if X and Z are two random variables, we may want to determine what is the probability that X takes some value x while at the same time Z takes some value z. If X and Z are discrete, they take only a countable number of values so that it is possible to assign probabilities directly to each pair of values (x,z). This yields the concept of joint probability mass function, which generalizes the concept of probability mass function.

definition 2.9: let X and Z be two discrete random variables; then X and Z have a **joint probability mass function** $f(x,z)$ such that $f(x,z) = \mathbb{P}(X = x, Z = z)$.

Consider again the usual 6-face die example:

example 2.1 (continued): let X be defined as $X(\omega) = 1$ if $\omega = 2,4,6$, and $X(\omega) = 0$ otherwise. Let $Z(\omega) = \omega$. The joint probability mass function $f(x,z)$ is then given by:

	$z=1$	$z=2$	$z=3$	$z=4$	$z=5$	$z=6$
$x=0$	$\mathbb{P}(\{1\})=1/6$	$\mathbb{P}(\varnothing)=0$	$\mathbb{P}(\{3\})=1/6$	$\mathbb{P}(\varnothing)=0$	$\mathbb{P}(\{5\})=1/6$	$\mathbb{P}(\varnothing)=0$
$x=1$	$\mathbb{P}(\varnothing)=0$	$\mathbb{P}(\{2\})=1/6$	$\mathbb{P}(\varnothing)=0$	$\mathbb{P}(\{4\})=1/6$	$\mathbb{P}(\varnothing)=0$	$\mathbb{P}(\{6\})=1/6$

Table 2.1: Joint probability mass function of X and Y

If instead X and Z are continuous, probabilities become defined by the joint probability density function, the generalisation of the density function.

definition 2.10: let X and Z be two continuous random variables; then X and Z have a **joint probability density function** $f(x,z)$ such that $\mathbb{P}(a \leq X \leq b, c \leq Z \leq d) = \int_a^b \int_c^d f(x,z)dzdx$.

Interestingly, it is possible to recover the probability functions of the individual random variables from their joint probability function. This is known as marginalisation.

definition 2.11: let X and Z be two random variables; the **marginal** probability mass or density function $f(x)$ obtains from:

$$f(x) = \sum_z f(x,z) \quad (X \text{ discrete}) \qquad \text{or} \qquad f(x) = \int_{-\infty}^{\infty} f(x,z)dz \quad (X \text{ continuous})$$

In other words, the marginal is obtained by summing over all the possible values of the other variable. This is illustrated by our usual example.

example 2.1 (continued): let X be defined as $X(\omega) = 1$ if $\omega = 2,4,6$, and $X(\omega) = 0$ otherwise. Let $Z(\omega) = \omega$. The marginal distributions $f(x)$ and $f(z)$ obtain from the joint mass function, as shown in Table 2.2:

	$z=1$	$z=2$	$z=3$	$z=4$	$z=5$	$z=6$	Marginal: $f(x)$
$x=0$	$\mathbb{P}(\{1\}) = 1/6$	$\mathbb{P}(\varnothing) = 0$	$\mathbb{P}(\{3\}) = 1/6$	$\mathbb{P}(\varnothing) = 0$	$\mathbb{P}(\{5\}) = 1/6$	$\mathbb{P}(\varnothing) = 0$	3/6
$x=1$	$\mathbb{P}(\varnothing) = 0$	$\mathbb{P}(\{2\}) = 1/6$	$\mathbb{P}(\varnothing) = 0$	$\mathbb{P}(\{4\}) = 1/6$	$\mathbb{P}(\varnothing) = 0$	$\mathbb{P}(\{6\}) = 1/6$	3/6
Marginal: $f(z)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	

Table 2.2: Marginal mass functions of X and Y

Section 2.2 introduced the concept of conditional probabilities for events. We now want to generalize the concept to random variables, with a similar interpretation. For instance, given two random variables X and Z, what is the probability that X takes some value x if we observe that Z has taken a given value z? This notion of conditional distribution is central in Bayesian analysis, and constitutes the foundation of Bayes law for random variables.

definition 2.12: let X and Z be two random variables; let $f(x,z)$, $f(x)$ and $f(z)$ respectively denote their joint and marginal probability mass (or density) functions. The **conditional probability mass function** (or **conditional probability density function**) is given by:

$$f(x|z) = \frac{f(x,z)}{f(z)}$$

Note the similarities between definition 2.12 and definition 2.3 in the case of events. To illustrate the concept, consider the usual 6-face die example:

example 2.1 (continued): let X be defined as $X(\omega) = 1$ if $\omega = 2,4,6$, and $X(\omega) = 0$ otherwise. Let $Z(\omega) = \omega$. Consider the difference between $f(x)$ and $f(x|z)$. For $x = 1$ and $z = 2$, Table 2.2 gives $f(x) = 3/6$, $f(z) = 1/6$ and $f(x,z) = 1/6$. So $f(x|z) = (1/6)/(1/6) = 1$. In other words, the unconditional probability $f(x)$ to observe $X = 1$ (the outcome is an even number) is 1/2. However, once $Z = 2$ is observed (the outcome of the roll is 2), we now for sure that the outcome is even and we update the probability to $f(x|z) = 1$.

A concept related to the idea of conditional distribution is that of independence. Informally, we say that two random variables X and Z are independent if "knowing Z tells us nothing about the value of X". Note that definition 2.12 implies that $f(x,z) = f(x|z)f(z)$. The intuition is then that if Z says nothing about X, the conditional density $f(x|z)$ should be equal to the unconditional density $f(x)$. This in turn yields $f(x,z) = f(x)f(z)$. In other words, when two random variables are independent, their joint density is just the product of the marginal densities.

definition 2.13: let X and Z be two random variables; X and Z are **independent** if for any x and z:

$$f(x,z) = f(x)f(z)$$

It is now possible to introduce the final and main result of this chapter. It follows directly from definition 2.11 that $f(x|z) = \dfrac{f(x,z)}{f(z)}$ and $f(x,z) = f(z|x)f(x)$. Substituting the latter in the former yields Bayes rule for random variables.

definition 2.14: let X and Z be two random variables; Bayes rule is given by:
$f(x|z) = \dfrac{f(z|x)f(x)}{f(z)}$

This simple formula constitutes the core of Bayesian analysis and will be used thoughout the whole book. Note again the similarities with Bayes rule for events given by definition 2.4. The formula says that the conditional density $f(x|z)$ is equal to the unconditional density $f(x)$, updated by the additional information $f(z|x)/f(y)$ obtained from the observation of Z.

example 2.1 (continued): let X be defined as $X(\omega) = 1$ if $\omega = 2,4,6$, and $X(\omega) = 0$ otherwise. Let $Z(\omega) = \omega$. For $x = 1$ and $z = 2$, Table 2.2 gives $f(x) = 3/6$, $f(z) = 1/6$, $f(x,z) = 1/6$, so that $f(y|z) = (1/6)/(3/6) = 1/3$.
Following, $f(x|z) = f(z|x)f(x)/f(z) = (1/3)(3/6)/(1/6) = 1$.
The unconditional density $f(x) = 1/2$ has been updated to $f(x|z) = 1$ once the value $Z = z$ has been observed.

CHAPTER 3

Three applied examples

Chapter 1 introduced the fundamental concepts of Bayesian statistics, while chapter 2 developed the technical framework leading to Bayes rule. This chapter puts these elements together and conducts the first actual applications of Bayesian statistics, building on the simple examples introduced in chapter 1 (a coin flip, the number of cars sold in a day, and the return of a stock at the New York Stock Exchange).

3.1 Principles of estimation

Recall from chapter 1 that our objective consists in estimating some parameter θ, using a sample of observations y. Under a frequentist approach, estimation by maximum likelihood is straightforward: obtain first the likelihood function $f(y|\theta)$ from the data, then find the value $\hat{\theta}$ that maximizes it.

In a Bayesian context however, estimation is conducted with Bayes rule. Definition 2.14 provides the general formula $f(x|z) = f(z|x)f(x)/f(z)$, for any two random variables X and Z. Since the Bayesian approach treats both the data y and the parameters θ as random variables, we can substitute for $x = \theta$ and $z = y$ to obtain the version of Bayes rule used in empirical applications.

definition 3.1: let y denote the sample of observations and θ the parameters of interest to estimate; **Bayes rule** is given by:

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

In the above definition, the use of $\pi(\theta|y)$ and $\pi(\theta)$ in place of $f(\theta|y)$ and $f(\theta)$ is a pure matter of notation. It is useful to take a closer look at the elements of definition 3.1.

On the left-hand side, $\pi(\theta|y)$ is the **posterior distribution**. It represents the distribution of the random variable θ, conditioned on having observed the data y, and represents the main object of interest.

On the right-hand side, $f(y|\theta)$ is the **likelihood function**. It is the density function of the observed data y for a given value of θ. It represents the information contained in the sample of observations.

The third term is the **prior distribution** $\pi(\theta)$ representing the subjective prior belief about θ. It constitutes the information available before the data is observed.

The final term is the **marginal likelihood** $f(y)$. It represents the unconditional density of the data, or in other words the data likelihood regardless of the value of θ. Often, this term cannot be estimated directly.

Definition 3.1 says that the posterior distribution $\pi(\theta|y)$ is equal to the prior distribution $\pi(\theta)$, updated by the additional information obtained from observing the data $f(y|\theta)$ and the overall data likelihood $f(y)$. If the marginal likelihood was known, Bayes rule 3.1 could be applied directly. In practice however this term is unknown, which motivates a brief but important digression.

Notice that the marginal likelihood $f(y)$ does not involve θ. In this respect, it only plays the role of a normalization constant ensuring that the posterior $\pi(\theta|y)$ integrates to 1, and carries no information on θ. It is then convenient to ignore it, using the notion of kernel.

> **definition 3.2:** let $f(x)$ be some probability density function that can be expressed as $f(x) = \alpha.g(x)$, with α a multiplicative term not involving x. Then we write:
> $f(x) \propto g(x)$
> which reads "$f(x)$ is proportional to $g(x)$". $g(x)$ is called the **kernel** of the density function $f(x)$, and α is called the **normalization constant** .

Definition 3.2 says that $f(x)$ is proportional to $g(x)$ up to some multiplicative constant α that only serves as a normalization device. In Bayesian analysis it is convenient to work with kernels rather than with the actual density functions, typically ignoring the normalization constant. For our purpose, an immediate application of this strategy consists in rewriting Bayes rule in definition 3.1 as a kernel to get rid of the marginal likelihood.

> **definition 3.3:** let y denote the sample of observations and θ the parameters of interest to estimate; **Bayes rule** is given by:
> $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$

Definition 3.3 says that the posterior $\pi(\theta|y)$ is proportional to the likelihood function $f(y|\theta)$ multiplied by the prior $\pi(\theta)$, up to the marginal likelihood $f(y)$ that represents the normalization constant and is ignored. The Bayesian estimation process then reduces to a trivial product between the likelihood function $f(y|\theta)$ and the prior $\pi(\theta)$.

We can now summarize the estimation procedures under the frequenctist and Bayesian approaches.

Summary of estimation procedures:

frequentist approach (maximum likelihood):

- set the likelihood function $f(y|\theta)$
- find $\hat{\theta}$ that maximizes $f(y|\theta)$

Bayesian approach (Bayes rule):

- set the likelihood function $f(y|\theta)$
- set the prior distribution $\pi(\theta)$
- apply $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$

3.2 A first example: flipping a coin

Consider again the coin flip example developed in chapter 1. Assume you want to determine the probability that a coin will come up with "heads". A simple strategy consists in flipping the coin n times, and observe the number m of "heads" outcomes.

The first step consists in setting a statistical model for the experiment. A simple choice consists in modelling each of the n flips as a Bernoulli distribution with probability of success p. The parameter of interest of the model is thus $\theta = \{p\}$. Denoting then by y_i the outcome of the i^{th} flip (1 for a success, 0 for a failure), the probability mass function for each flip is given by:

$$f(y_i|p) = p^{y_i}(1-p)^{1-y_i} \tag{1.3.1}$$

Start with a frequentist estimate of θ. Following the procedure suggested in section 3.2, we need to set the likelihood function $f(y|\theta)$, which represents the density function for the sample of observations as a whole. Equation (1.3.1) only provides the density for a single observation. To obtain the joint density over the whole sample, we assume that the observations are generated independently. Then from definition 2.13, the joint density becomes the product of the individual densities.

definition 3.4: let $y = y_1, y_2, \cdots, y_n$ denote a sample of n observations, with $f(y_i|\theta)$ the density of each individual observation. The **likelihood function** $f(y|\theta)$ obtains by assuming independence between the observations, so that:

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Applying definition 3.4 to the individual densities (1.3.1), the likelihood function obtains as:

$$f(y|p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \tag{1.3.2}$$

After some manipulations, it can be shown (book 2, p. 3) that the likelihood function rewrites as:

$$f(y|p) = p^m(1-p)^{n-m} \tag{1.3.3}$$

A maximum likelihood estimate can then be obtained by finding the value $\hat{\theta}$ that maximizes the likelihood function $f(y|p)$. In practice, it is often easier to maximize the logarithm of the likelihood. This is equivalent since extrema are not affected by monotonic transformations.

definition 3.5: let $f(y|\theta)$ denote the likelihood function; the **maximum likelihood estimate** $\hat{\theta}$ obtains by maximizing the log-likelihood function:

$$\hat{\theta} = \underset{\theta}{argmax} \;\; log(f(y|\theta))$$

Taking the log of the likelihood function (1.3.3), the maximum likelihood estimate becomes:

$$\hat{p} = \underset{p}{argmax} \;\; m\, log(p) + (n-m)\; log(1-p) \tag{1.3.4}$$

The maximum is found by setting the derivative with respect to p to 0 and solving for p (book 2, p. 3). This yields:

$$\hat{p} = m/n \tag{1.3.5}$$

The maximum likelihood estimate $\hat{p}$ is thus the proportion of observed successes over the total number of trials, or in other words the empirical mean.

Consider now a Bayesian estimate of p. The procedure developed in section 3.2 first requires the likelihood function $f(y|p)$, which is already known (equation (1.3.3)). We then need a prior distribution $\pi(p)$ for p. Since p represents a probability, we want a prior distribution that produces values between 0 and 1. The Beta distribution then constitutes a good candidate. Its density is given by:

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{1.3.6}$$

α and β are constants that determine the overall shape of the distribution. They are known as hyperparameters.

definition 3.6: a **hyperparameter** is a parameter which defines the prior distribution.

The choice of values for α and β will be discussed shortly. For now, we implement the final step of the estimation procedure, applying Bayes rule 3.3 to the likelihood function (1.3.3) and the prior distribution (1.3.6). This yields:

$$\pi(p|y) \propto p^m(1-p)^{n-m} \times p^{\alpha-1}(1-p)^{\beta-1} \tag{1.3.7}$$

Notice that the multiplicative constant $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ in equation (1.3.6) has been dropped. This is because it does not involve p, and can hence also be relegated to the normalization constant when working with the kernel of the posterior $\pi(p|y)$. Gathering the terms in (1.3.7), we obtain:

$$\pi(p|y) \propto p^{\bar{\alpha}-1}(1-p)^{\bar{\beta}-1} \tag{1.3.8}$$

with:

$$\bar{\alpha} = \alpha + m \qquad \bar{\beta} = \beta + n - m \tag{1.3.9}$$

Looking at equation (1.3.8), we recognize the kernel of a Beta distribution with shape parameters $\bar{\alpha}$ and $\bar{\beta}$. Following, we conclude that the posterior distribution is Beta with shapes $\bar{\alpha}$ and $\bar{\beta}$: $\pi(p|y) \sim Beta(\bar{\alpha}, \bar{\beta})$. Interestingly, the posterior distribution belongs to the same family as the prior: this is known as a conjugate distribution.

definition 3.7: a prior and a posterior distribution are called **conjugate distributions** if they belong to the same family of distribution.

Let us now consider a numerical example. Assume the coin is flipped $n = 100$ times, and yields heads $m = 63$ times. The maximum likelihood estimate for p is thus $\hat{p} = m/n = 63/100$ or 0.63.

Consider now the Bayesian estimate. We first need to set the values of α and β for the prior $\pi(p)$. The choice must reflects our personal belief about the distribution, and will have a significant impact on the posterior distribution. Figure 3.1 shows the Beta density functions for different values of α and β.

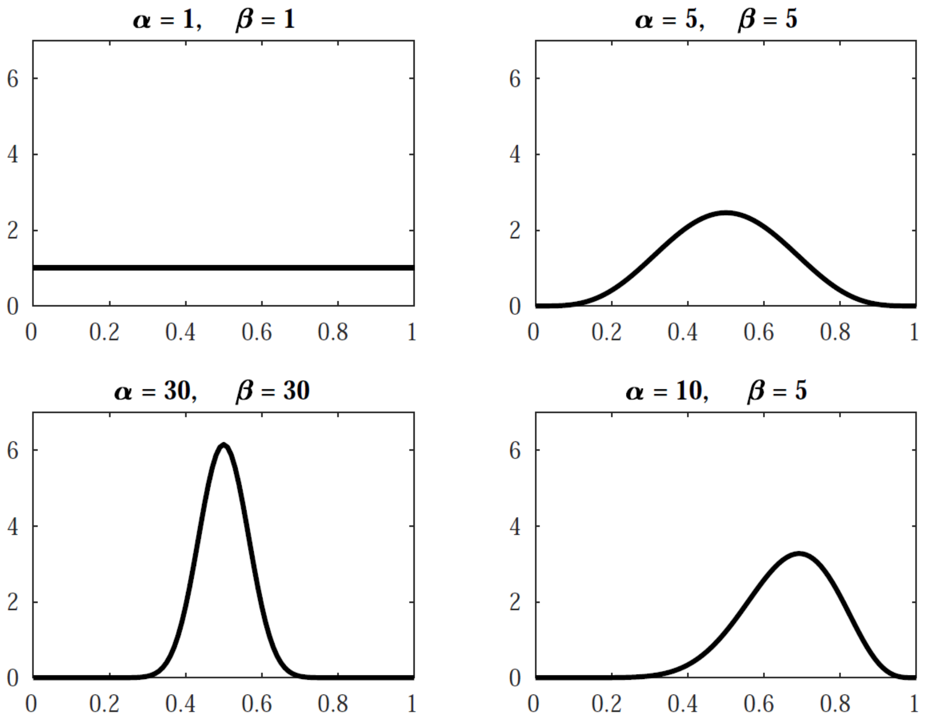


Figure 3.1: Probability density function of the Beta distribution for different α and β values

It can be seen that the distribution is symmetric around 0.5 for $\alpha = \beta$, and skewed otherwise. Also, the larger α and β, the tighter the distribution and the smaller the variance. So, what could be good values of α and β? Things get really subjective here, but the following propositions are reasonable. First, coins should be balanced on average, with the same chance to biased upward or downward. This implies a symmetric distribution centered at 0.5, and thus $\alpha = \beta$. Also, a potential bias should be reasonably small. Assuming for instance that the typical probability of success is comprised between 0.45 and 0.55 yields a standard deviation of 0.05, and from property d.26 of the Beta distribution this is obtained by setting $\alpha = \beta = 40$. Given these choices for the prior, we can eventually calculate the posterior parameters: $\bar{\alpha} = \alpha + m = 40 + 63 = 103$ and $\bar{\beta} = \beta + n - m = 40 + 100 - 63 = 77$.

The whole example is represented on Figure 3.2. The dashed line on the right is the likelihood function, peaking at the maximum likelihood estimate $\hat{p} = 0.63$. The left grey curve represents the prior distribution. As implied by our choice for α and β, it is symmetric around its mean of 0.5, and has a 0.05 standard deviation. Finally, the black plain line in the middle reflects the posterior distribution. It appears as a compromise between the prior and the likelihood, with a mean of approximately 0.57, somewhere between the prior mean and the maximum likelihood estimate.

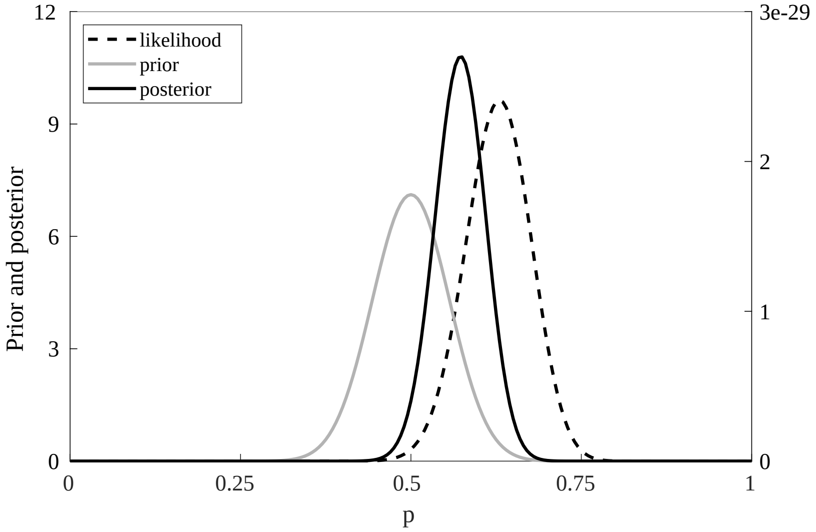


Figure 3.2: Likelihood, prior and posterior for the coin flip example

3.3 A second example: modelling monthly car sales

Consider now the car sales example developed in chapter 1. A car retailer is interested in predicting the monthly sales of a local outlet store to check how profitable the store is. To do so, a history of n month is collected with the observed sales for each month.

We first set a statistical model for the experiment. Because the monthly sales are some integer between 0 and infinity, a simple choice is a Poisson model with an intensity of λ. The parameter of interest is thus $\theta = \{\lambda\}$. Denoting by y_i the sales of month i, the probability mass fonction for each month is given by:

$$f(y_i|\lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \tag{1.3.10}$$

Consider first a frequentist estimate of θ. Following the procedure suggested in section 3.2, we first need the likelihood function $f(y|\theta)$. Using the individual densities (1.3.10) and definition 3.4, it can be shown (book 2, p. 3) that the likelihood function obtains as:

$$f(y|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} y_i}\, e^{-n\lambda}}{\prod_{i=1}^{n} y_i!} \tag{1.3.11}$$

Applying definition 3.5 and taking the log of the likelihood function (1.3.11), a maximum likelihood estimate of λ obtains from:

$$\hat{\lambda} = \underset{\lambda}{argmax} \left(\sum_{i=1}^{n} y_i\right) log(\lambda) - n\lambda - \sum_{i=1}^{n} log(y_i!) \tag{1.3.12}$$

The maximum is found by setting the derivative with respect to λ to 0 and solving for λ (book 2, p. 3):

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.3.13}$$

The maximum likelihood estimate $\hat{\lambda}$ is thus simply the empirical mean over the sample of observations.

Consider now a Bayesian estimate of λ. The likelihood function $f(y|\lambda)$ is already known (equation (1.3.11)). We then need a prior distribution $\pi(\lambda)$ for λ. Since λ represents both the mean and variance of the Poisson distribution, we need a prior that produces positive values. The Gamma distribution then represents a good candidate. Its density is given by:

$$\pi(\lambda) = \frac{b^{-a}}{\Gamma(a)}\, \lambda^{a-1}\, e^{-\lambda/b} \tag{1.3.14}$$

a and b are the shape and scale hyperparameters of the Gamma distribution whose values will be discussed shortly. For now, we apply Bayes rule 3.3 to the likelihood function (1.3.11) and the prior distribution (1.3.14) to obtain:

$$\pi(\lambda|y) \propto \lambda^{\sum_{i=1}^{n} y_i}\, e^{-n\lambda} \times \lambda^{a-1}\, e^{-\lambda/b} \tag{1.3.15}$$

Again, all the multiplicative terms not involving λ have been relegated to the normalization constants. Rearranging yields (book 2, p. 4):

$$\pi(\lambda|y) \propto \lambda^{\bar{a}-1}\, e^{-\lambda/\bar{b}} \tag{1.3.16}$$

with:

$$\bar{a} = a + \sum_{i=1}^{n} y_i \qquad \bar{b} = \frac{b}{bn+1} \tag{1.3.17}$$

Looking at equation (1.3.16), we recognize the kernel of a Gamma distribution with shape $\bar{a}$ and scale $\bar{b}$. Following, we conclude that $\pi(\lambda|y) \sim G(\bar{a},\bar{b})$. Again, we have here an example of a conjugate distribution.

Let us now consider a numerical example. Assume the retailer has an history of 5 years of monthly sales for the store, i.e., a sample of 60 observations. The total sales over the sample is 505, for a sample mean of 8.42. The maximum likelihood estimate for λ is thus $\hat{\lambda} = 8.42$.

Consider now the Bayesian estimate. We first need to set the values of a and b for the prior $\pi(\lambda)$. Assume the retailer knows from the data records of other stores in the district that the average monthly sales of cars are 11.2, with a variance of 0.16. The prior belief is thus a Gamma distribution with a mean of 11.2 and a variance of 0.16. From property d.19 of the Gamma distribution, this can be achieved by setting $a = 784$ and $b = 0.0143$. Given these choices for the prior, we can eventually calculate the posterior parameters: $\bar{a} = a + \sum_{i=1}^{n} y_i = 784 + 505 = 1289$ and $\bar{b} = \frac{b}{bn+1} = 0.0077$, implying a posterior mean of 9.92.

The whole example is represented on Figure 3.3.

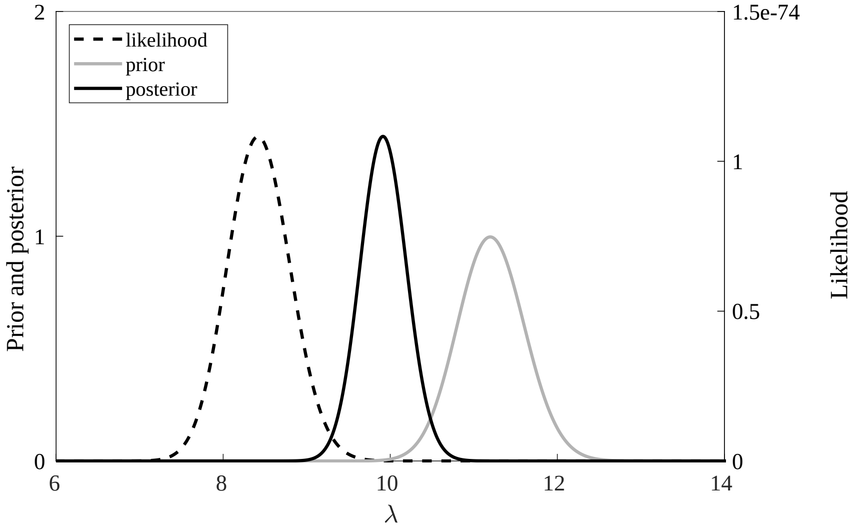


Figure 3.3: Likelihood, prior and posterior for the car sales example

The likelihood function is depicted by the left dashed line, peaking at the maximum likelihood estimate of 8.42. The grey line on the right represents the Gamma prior with the constructed mean of 11.2 and standard deviation of 0.4. In the middle, the black line shows the Gamma posterior with a mean of 9.92. Even though the car retailer had a prior opinion of an average 11.2 sales a month, the empirical evidence suggested a smaller value of 8.42. The final belief accounts for both sources of information and lies somewhere in-between, at an average of 9.92.

3.4 A third example: predicting a stock return

Consider finally the third example introduced in chapter 1. An investor wants to predict the return of a given stock traded on the NYSE. To do so, a sample of n past annual return values is collected for the stock.

We first set a statistical model for the experiment. Because returns can take any positive or negative values, a normal distribution constitutes a good candidate. This distribution is characterized by a mean parameter μ and a variance parameter σ, which respectively represent the average return and the volatility of the stock. For now we keep things simple and assume that the stock volatility σ is known. The only parameter remaining to estimate for the investor is thus the average return μ, so that $\theta = \{\mu\}$. Denoting by y_i the stock return on year i, the probability density function for each return is given by:

$$f(y_i|\mu) = (2\pi\sigma)^{-1/2} \exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma}\right) \tag{1.3.18}$$

Consider first a frequentist estimate of θ. Following the procedure suggested in section 3.2, we first set the likelihood function $f(y|\theta)$. Using the individual densities (1.3.18) and definition 3.4, it can be shown (book 2, p. 4) that the likelihood function obtains as:

$$f(y|\mu) = (2\pi\sigma)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma}\right) \tag{1.3.19}$$

Applying definition 3.5 and taking the log of the likelihood function (1.3.19), a maximum likelihood estimate of μ obtains from:

$$\hat{\mu} = \underset{\mu}{argmax} \quad -n/2 \; log(2\pi\sigma) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma} \tag{1.3.20}$$

The maximum is found by setting the derivative with respect to μ to 0 and solving for μ (book 2, p. 4):

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.3.21}$$

The maximum likelihood estimate $\hat{\mu}$ is thus simply the empirical mean over the sample of observations.

Consider now a Bayesian estimate of μ. The likelihood function $f(y|\lambda)$ is already known (equation (1.3.19)). We then set a prior distribution $\pi(\mu)$ for μ. Since μ represents the average stock return, it can take any real value. The normal distribution thus represents a good candidate, with a density given by:

$$\pi(\mu) = (2\pi v)^{-1/2} \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v}\right) \tag{1.3.22}$$

m and v are hyperparameters respectively representing the mean and variance of the prior distribution. Next, we apply Bayes rule 3.3 to the likelihood function (1.3.19) and the prior distribution (1.3.22). This yields:

$$\pi(\mu|y) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma}\right) \times \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v}\right) \tag{1.3.23}$$

Again, any multiplicative term not involving μ has been relegated to the normalization constants. Intuitively, because (1.3.23) involves two normal distributions, the posterior should be normal as well. The difficulty consists in turning the pair of normal densities into a single one, and the methodology to do so is known as completing the squares.

definition 3.8: completing the squares is the methodology combining a normal likelihood function $f(y|\theta)$ with a normal prior $\pi(\theta)$ to obtain a normal posterior $\pi(\theta|y)$.

Completing the squares is used again and again throughout the book, so it is useful to detail it step by step. First start from (1.3.23), develop the quadratic forms and group the terms to obtain (book 2, p. 5):

$$\pi(\mu|y) \propto \exp\left(-\frac{1}{2}\left[\mu^2\left(\frac{n}{\sigma}+\frac{1}{v}\right) - 2\mu\left(\frac{1}{\sigma}\sum_{i=1}^{n}y_i + \frac{m}{v}\right) + \frac{1}{\sigma}\sum_{i=1}^{n}y_i^2 + \frac{m^2}{v}\right]\right) \tag{1.3.24}$$

To complete the squares, we then add terms in (1.3.24) to make it factorable into a single quadratic form.

$$\pi(\mu|y) \propto \exp\left(-\frac{1}{2}\left[\mu^2\left(\frac{n}{\sigma}+\frac{1}{v}\right) - 2\mu\frac{\bar{v}}{\bar{v}}\left(\frac{1}{\sigma}\sum_{i=1}^{n}y_i + \frac{m}{v}\right) + \frac{1}{\sigma}\sum_{i=1}^{n}y_i^2 + \frac{m^2}{v} + \frac{\bar{m}^2}{\bar{v}} - \frac{\bar{m}^2}{\bar{v}}\right]\right) \tag{1.3.25}$$

We have multiplied the second term by $\bar{v}/\bar{v}$, and added and subtracted the quadratic term $\bar{m}^2/\bar{v}$. Clearly, (1.3.24) and (1.3.25) are equal, whatever the definition we choose for $\bar{m}$ and $\bar{v}$. The trick however consists in finding the right definition to permit factorization. The values we want are:

$$\bar{v} = \left(\frac{n}{\sigma}+\frac{1}{v}\right)^{-1} \qquad \bar{m} = \bar{v}\left(\frac{1}{\sigma}\sum_{i=1}^{n}y_i + \frac{m}{v}\right) \tag{1.3.26}$$

Substituting this back in (1.3.25) eventually yields:

$$\pi(\mu|y) \propto \exp\left(-\frac{1}{2}\left[\frac{\mu^2}{\bar{v}} - 2\mu\frac{\bar{m}}{\bar{v}} + \frac{\bar{m}^2}{\bar{v}} + \frac{1}{\sigma}\sum_{i=1}^{n}y_i^2 + \frac{m^2}{v} - \frac{\bar{m}^2}{\bar{v}}\right]\right) \tag{1.3.27}$$

Now we can factor the first three terms into a single quadratic form that will be the kernel of the posterior, and set the final three terms as a separate multiplicative constant:

$$\pi(\mu|y) \propto \exp\left(-\frac{1}{2}\frac{(\mu-\bar{m})^2}{\bar{v}}\right)\exp\left(-\frac{1}{2}\left[\frac{1}{\sigma}\sum_{i=1}^{n}y_i^2+\frac{m^2}{v}-\frac{\bar{m}^2}{\bar{v}}\right]\right) \tag{1.3.28}$$

Noting finally that the second multiplicative term does not involve μ, it can be relegated to the normalization constant to yield:

$$\pi(\mu|y) \propto \exp\left(-\frac{1}{2}\frac{(\mu-\bar{m})^2}{\bar{v}}\right) \tag{1.3.29}$$

We eventually recognize in (1.3.29) the kernel of a normal distribution, and conclude that the posterior distribution of μ is normal with mean $\bar{m}$ and variance $\bar{v}$: $\pi(\mu|y) \sim N(\bar{m},\bar{v})$. We have succesfully applied the completing the squares methodology, constituted of equations (1.3.23) - (1.3.29). Also, we note that we face again a case of conjugate distributions since both the prior and the posterior are normal.

Let us now consider a numerical example. Assume the investor has a history of 20 years of yearly returns on the stock, i.e., a sample of $n = 20$ observations. The mean annual return over the sample sample mean is \$18.2, with a known variance of $\sigma = 5.2$. The maximum likelihood for the average return is thus $\hat{\mu} = 18.2$.

Consider now the Bayesian estimate. We first set the values of m and v for the prior $\pi(\mu)$. Assume the investor has made his calculations about the future profits of the company and expects an average annual return of \$12.7 with a variance of 0.4. The prior belief is thus $m = 12.7$ and $v = 0.4$. It is then possible to calculate the posterior parameters using equation (1.3.28), yielding $\bar{m} = 16.03$ and $\bar{v} = 0.16$.

The whole example is represented on Figure 3.4.

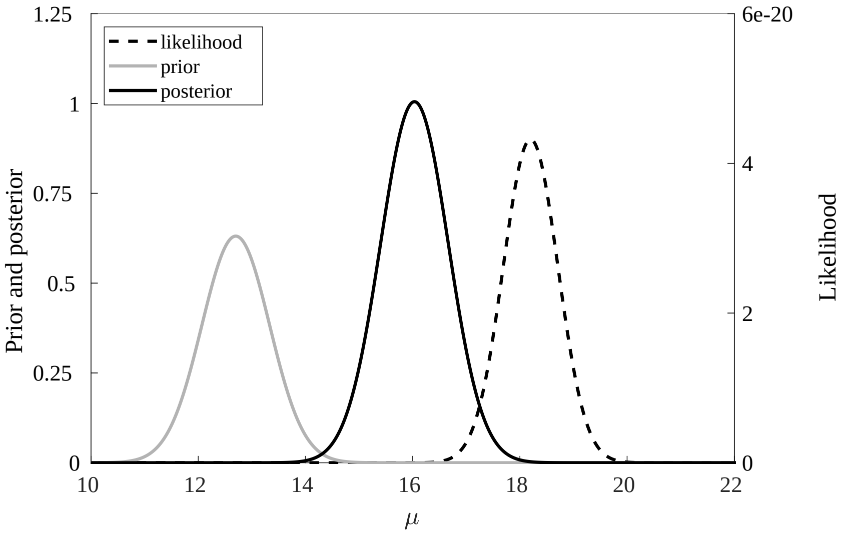


Figure 3.4: Likelihood, prior and posterior for the stock return example

The likelihood function is depicted by the right dashed line, peaking at the maximum likelihood estimate of 18.2. The grey line on the left represents the normal prior with a mean of 12.7 and a variance of 0.4. In the middle, the black line shows the normal posterior with a mean of 16.03. The investor had a prior opinion of an average stock return of \$12.7. Yet, empirical evidence suggested a much better performance of \$18.2. The final posterior belief represents a compromise, with an updated average return of \$16.03.

CHAPTER 4

Further aspects of Bayesian priors and posteriors

Chapter 3 introduced the fundamentals of Bayesian analysis with three simple examples. In this chapter we develop a number of additional aspects of Bayesian models that arise in practical applications.

4.1 Multivariate priors

Most practical Bayesian applications involve several parameters, so that $\theta = \{\theta_1, \cdots, \theta_n\}$. In this case, Bayes rule is still given by definition 3.3 as $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$, but the prior $\pi(\theta) = \pi(\theta_1, \cdots, \theta_n)$ and the posterior $\pi(\theta|y) = \pi(\theta_1, \cdots, \theta_n|y)$ now denote joint densities.

To define a joint prior, one simply assumes independence between the different parameters $\theta_1, \cdots, \theta_n$ so that from definition 2.13, the joint prior is the product of the individual priors.

definition 4.1: let $\theta = \{\theta_1, \cdots, \theta_n\}$ be the model parameters; the **joint prior distribution** is obtained by assuming independence between the parameters, so that:
$\pi(\theta_1, \cdots, \theta_n) = \pi(\theta_1) \ \cdots \ \pi(\theta_n)$

To illustrate this, consider again the stock return example developed in chapter 3:

example 4.1: an investor wants to predict the return on a given stock. The statistical model for the stock return is a normal distribution with mean μ and variance σ. We now assume that both μ and σ are unknown, hence the parameters of interest to estimate are $\theta = \{\mu, \sigma\}$.

Following definition 3.3, Bayes rule for the model is $\pi(\mu, \sigma|y) \propto f(y|\mu, \sigma)\pi(\mu, \sigma)$. Given definition 4.1 we assume independence between μ and σ so that $\pi(\mu, \sigma|y) \propto f(y|\mu, \sigma)\pi(\mu)\pi(\sigma)$.

The likelihood $f(y|\mu, \sigma)$ and the prior $\pi(\mu)$ are already known and given by (1.3.19) and (1.3.22). Because σ represents a variance term, it takes only positive values. The inverse Gamma distribution is then a good choice and we set $\pi(\sigma) \sim IG(\alpha/2, \delta/2)$, where α and δ respectively denote the shape and scale hyperparameters of the distribution[1]. The prior density is then:

$$\pi(\sigma) = \frac{\delta/2^{\alpha/2}}{\Gamma(\alpha/2)} \sigma^{-\alpha/2-1} \exp\left(-\frac{\delta}{2\sigma}\right) \tag{1.4.1}$$

Applying Bayes rule $\pi(\mu, \sigma|y) \propto f(y|\mu, \sigma)\pi(\mu)\pi(\sigma)$ and relegating to the normalization constant any term not involving μ or σ, we eventually obtain the kernel of the joint posterior as:

[1] The inverse Gamma is here preferred over the Gamma distribution. This is because the inverse Gamma is conjugate with the normal likelihood, while the Gamma is not and hence does not yield tractable posteriors. The division of the hyperparameters α and δ by 2 is also for conjugacy with the normal likelihood.

$$\pi(\mu,\sigma|y) \propto \sigma^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma}\right) \times \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v}\right) \times \sigma^{-\alpha/2-1} \exp\left(-\frac{\delta}{2\sigma}\right) \tag{1.4.2}$$

What to do with this joint posterior will be the subject of section 4.3.

4.2 Hierarchical priors

We have seen in definition 3.6 that prior distributions are defined by parameters called hyperparameters. Most of the time, these hyperparameters are constants exogenously supplied by the statistician. Sometimes however we want to add one level to the model by assuming that the hyperparameters themselves are random variables which are assigned a prior distribution and integrated in the estimation process.

definition 4.2: let θ be a parameter whose prior distribution is conditional on some hyperparameter λ; a **hierarchical prior** is a prior which considers λ as a random variable and assigns it a prior distribution $\pi(\lambda)$, known as a **hyperprior**.

Because the hyperparameter λ is treated as a random variable, the prior $\pi(\theta)$ becomes a joint prior $\pi(\theta,\lambda)$. From definition 2.12, this joint prior can then rewrite as $\pi(\theta,\lambda) = \pi(\theta,\lambda)/\pi(\lambda) \times \pi(\lambda) = \pi(\theta|\lambda)\pi(\lambda)$. In other words, the hierarchical prior is expressed as a product of the conditional prior $\pi(\theta|\lambda)$ with the hyperprior $\pi(\lambda)$.

To illustrate this, consider again the stock return example.

example 4.1 (continued): we still model the stock return as a normal distribution with mean μ and variance σ. However, we set a hierarchical prior for μ by assuming that its prior variance depends on the stock volatility parameter σ. Precisely, we set $\pi(\mu|\sigma) \sim N(m, v\sigma)$, so that:

$$\pi(\mu|\sigma) = (2\pi v\sigma)^{-1/2} \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v\sigma}\right) \tag{1.4.3}$$

This prior is similar to (1.3.22) except that the variance is now also proportional to σ.

The two parameters of the model are $\theta = \{\mu,\sigma\}$, and Bayes rule is $\pi(\mu,\sigma|y) \propto f(y|\mu,\sigma)\pi(\mu,\sigma)$. Given the hierarchical prior, we rewrite $\pi(\mu,\sigma) = \pi(\mu|\sigma)\pi(\sigma)$ and Bayes rule becomes $\pi(\mu,\sigma|y) \propto f(y|\mu,\sigma)\pi(\mu|\sigma)\pi(\sigma)$. The likelihood $f(y|\mu,\sigma)$ and the hyperprior $\pi(\sigma)$ are given by (1.3.19) and (1.4.1). Combining with the prior $\pi(\mu|\sigma)$ given by (1.4.3) and relegating to the normalization constant any term not involving μ or σ, the kernel of the posterior then obtains as:

$$\pi(\mu,\sigma|y) \propto \sigma^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma}\right) \times \sigma^{-1/2} \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v\sigma}\right) \times \sigma^{-\alpha/2-1} \exp\left(-\frac{\delta}{2\sigma}\right) \tag{1.4.4}$$

4.3 Marginal posteriors

Most Bayesian models involve several parameters $\theta_1,\cdots,\theta_n$. In this case, Bayes rule yields a joint posterior distribution $\pi(\theta_1,\cdots,\theta_n|y)$. As such, the joint posterior is not interpretable. We thus want to derive the marginal posterior distributions $\pi(\theta_1|y),\cdots,\pi(\theta_n|y)$ for each individual parameter. This is done by marginalizing the joint posterior, as provided by definition 2.11.

definition 4.3: let $\pi(\theta_1, \cdots, \theta_n|y)$ be a joint distribution; the **marginal posterior distributions** $\pi(\theta_1|y), \cdots, \pi(\theta_n|y)$ obtain by integrating out the remaining parameters, so that:
$\pi(\theta_i|y) = \int \pi(\theta_1, \cdots, \theta_n|y) d\theta_{\neq i}$

Marginalization with definition 4.3 may or may not be possible, depending on the form of the posterior distribution. To see this, consider again the stock return example.

example 4.1 (continued): sections 4.1 and 4.2 both provide a joint posterior $\pi(\mu, \sigma|y)$ for the stock return example. Start with the hierarchical prior of section 4.2, which results in the posterior (1.4.4). It is possible to marginalize this posterior, though some work is required. First develop, group the terms and complete the squares to obtain:

$$\pi(\mu, \sigma|y) \propto \sigma^{-1/2} \exp\left(-\frac{1}{2}\frac{(\mu - \bar{m})^2}{\sigma\bar{v}}\right) \times \sigma^{-\bar{\alpha}/2-1} \exp\left(-\frac{\bar{\delta}}{2\sigma}\right) \tag{1.4.5}$$

with:

$$\bar{v} = \left(n + \frac{1}{v}\right)^{-1} \qquad \bar{m} = \bar{v}\left(\sum_{i=1}^{n} y_i + \frac{m}{v}\right) \qquad \bar{\alpha} = \alpha + n \qquad \bar{\delta} = \delta + \sum_{i=1}^{n} y_i^2 + \frac{m^2}{v} - \frac{\bar{m}^2}{\bar{v}} \tag{1.4.6}$$

This reformulation makes it easier to marginalize for μ and σ. We can see that (1.4.5) is a product of two kernels: the kernel of a normal distribution with mean $\bar{m}$ and variance $\bar{v}$, and the kernel of an inverse Gamma distribution with shape $\bar{\alpha}/2$ and scale $\bar{\delta}/2$.

We then obtain the marginal posterior distributions $\pi(\sigma|y)$ and $\pi(\mu|y)$ from direct application of definition 4.3. Calculations are easy for σ: since μ only appears in the first density as the kernel of a normal distribution, integration yields a constant, leaving only the second kernel:

$$\begin{aligned}\pi(\sigma|y) = \int \pi(\mu, \sigma|y) d\mu \;&\propto\; \sigma^{-\bar{\alpha}/2-1} \exp\left(-\frac{\bar{\delta}}{2\sigma}\right) \int \sigma^{-1/2} \exp\left(-\frac{1}{2}\frac{(\mu - \bar{m})^2}{\sigma\bar{v}}\right) d\mu \\ &\propto\; \sigma^{-\bar{\alpha}/2-1} \exp\left(-\frac{\bar{\delta}}{2\sigma}\right)\end{aligned} \tag{1.4.7}$$

We recognize the kernel of an inverse Gamma distribution with shape $\bar{\alpha}/2$ and scale $\bar{\delta}/2$: $\pi(\sigma|y) \sim IG(\bar{\alpha}/2, \bar{\delta}/2)$.

The calculations are trickier for μ. As σ appears in all the terms of (1.4.5), we group them and integrate:

$$\pi(\mu|y) = \int \pi(\mu, \sigma|y) d\sigma \propto \int \sigma^{-(\bar{\alpha}+1)/2-1} \exp\left(-\frac{\bar{\delta} + (\mu - \bar{m})^2/\bar{v}}{2\sigma}\right) d\sigma \tag{1.4.8}$$

Now, here is the trick: we recognize in (1.4.8) the kernel of an inverse Gamma distribution with shape $(\bar{\alpha}+1)/2$ and scale $(\bar{\delta} + (\mu - \bar{m})^2/\bar{v})\ /\ 2$. Now, from definition 2.8 of the probability density function and definition 3.2 of the kernel, one obtains $\int f(x)dx = \alpha \int g(x)dx = 1$ so that $\int g(x)dx = 1/\alpha$. In other words, integrating the kernel yields the reciprocal of the normalization constant of the distribution. Applied to the inverse Gamma kernel (1.4.8), this yields:

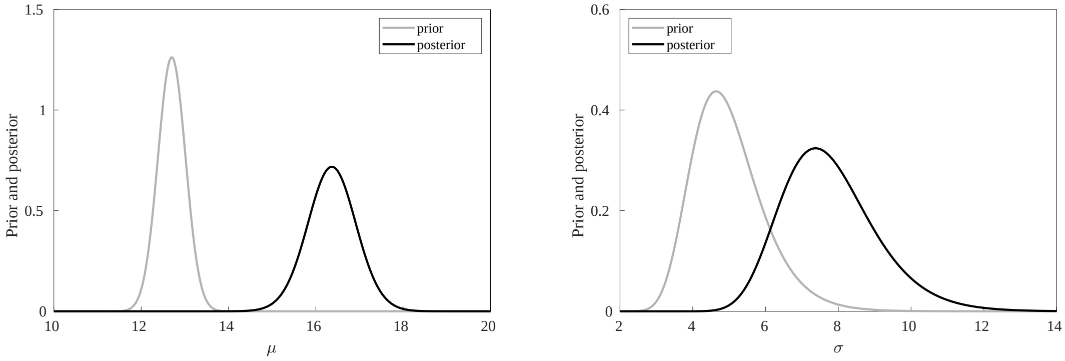
$$\pi(\mu|y) \propto \Gamma\left(\frac{\bar{\alpha}+1}{2}\right)\left(\frac{\bar{\delta} + (\mu - \bar{m})^2/\bar{v}}{2}\right)^{-\frac{\bar{\alpha}+1}{2}} \tag{1.4.9}$$

After some manipulations, it can be shown (book 2, p. 8) that this reformulates as :

$$\pi(\mu|y) \propto \left(1 + \frac{1}{\bar{\alpha}}\frac{(\mu - \bar{m})^2}{\bar{\delta}\bar{v}/\bar{\alpha}}\right)^{-\frac{\bar{\alpha}+1}{2}} \tag{1.4.10}$$

This is the kernel of a Student distribution with location $\bar{m}$, scale $\bar{\delta}\bar{v}/\bar{\alpha}$ and degrees of freedom $\bar{\alpha}$: $\pi(\mu|y) \sim T(\bar{m}, \bar{\delta}\bar{v}/\bar{\alpha}, \bar{\alpha})$.

We now continue the numerical example introduced in section 3.4. We keep the same values as before except for the prior variance on μ that is reduced to $v = 0.1$ to compensate for the additional uncertainty implied by the proportionality with σ in $\pi(\mu|\sigma) \sim N(m, v\sigma)$. Also, we need to define the hyperparameters α and δ for the prior $\pi(\sigma)$ defined in (1.4.1). Because the data suggests a variance around 5, we set an inverse Gamma distribution with a mean of 5 and a variance of 1. From property d.23 of the inverse Gamma distribution, this is obtained by setting $\alpha = 54$ and $\delta = 260$. We then obtain the posterior values $\bar{m} = 16.36$, $\bar{v} = 0.033$, $\bar{\alpha} = 74$ and $\bar{\delta} = 560.47$. The implied marginal posterior distributions along with the priors[2] are depicted in Figure 4.1:



(a) Prior and marginal posterior for μ

(b) Prior and marginal posterior for σ

Figure 4.1: Marginal posterior distributions for μ and σ

The Student marginal posterior for μ peaks at its average of 16.36, far from the prior distribution and its mean of 12.7, showing that much of the data evidence has been taken into account to update the prior belief. Similarly, the inverse Gamma marginal posterior for σ implies a mean of 7.78, implying a volatility larger than the prior belief of 5.

What if we now try to marginalize the joint posterior distribution (1.4.2) resulting from independent priors for μ and σ? It turns out that in this case marginalization using definition 4.3 is not possible. The terms involving μ and σ are too interwoven to calculate the integrals. In this case one must rely on simulation methods, which will be the object of part II of the book.

4.4 Point estimates

The posterior distribution $\pi(\theta|y)$ summarizes all the available information about θ. It thus constitutes the basis of any inference procedure. Suppose we want to obtain a single-value estimate of θ, based on the posterior distribution. The idea is to set a loss function $L(\hat{\theta}, \theta)$ which measures the loss incurred if the estimate is $\hat{\theta}$, but the true value is θ.

definition 4.4: let $\pi(\theta|y)$ denote the posterior distribution of some parameter θ; the point estimate of θ, called the **Bayes estimator** and denoted by $\hat{\theta}$ is the value that minimizes the expectation of some loss function:

$$\hat{\theta} = \underset{\theta}{argmin}\ \mathbb{E}[L(\hat{\theta}, \theta)] = \underset{\theta}{argmin} \int L(\hat{\theta}, \theta)\pi(\theta|y)d\theta$$

[2] Using $\sigma = 1$ for the hierarchical prior $\pi(\mu|\sigma)$ of μ.

Several different loss functions are possible. Classical choices are the quadratic loss function $L(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$, the absolute-value loss function $L(\hat{\theta},\theta) = |\hat{\theta}-\theta|$, and the 0-1 loss function $L(\hat{\theta},\theta) = \mathbb{1}(|\hat{\theta}-\theta| > c)$, with $\mathbb{1}(.)$ the indicator function and c some positive constant.

Consider for instance the quadratic loss function. From definition 4.4, the Bayes estimator obtains from:

$$\hat{\theta} = \underset{\theta}{argmin} \int (\hat{\theta}-\theta)^2 \pi(\theta|y) d\theta \tag{1.4.11}$$

The minimum is found by calculating the derivative of the function and setting it equal to zero, which yields:

$$2\int (\hat{\theta}-\theta)\pi(\theta|y)d\theta = 0 \tag{1.4.12}$$

Solving finally for $\hat{\theta}$ (book 2, p. 8), the Bayes estimator is given by:

$$\hat{\theta} = \int \theta \; \pi(\theta|y)d\theta \tag{1.4.13}$$

From (1.4.13) we conclude that $\hat{\theta} = \mathbb{E}(\theta|y)$: the Bayes estimator under the quadratic loss function is simply the mean of the posterior distribution $\pi(\theta|y)$. Alternative loss functions yield different point estimators, typically corresponding to some measure of central tendency. The absolute-value loss function for instance yields the median as the Bayes estimator, while the 0-1 loss function results in the mode when $c \to 0$. In practical applications the median is often prefered over the mean and the mode due to its robustness to extreme values.

example 4.1 (continued): consider point estimates for the parameters μ and σ in the stock return example, using the marginal posteriors developed in section 4.3. We retain the median as a point estimate. For μ, the marginal posterior is $\pi(\mu|y) \sim T(\bar{m}, \bar{\delta}\bar{v}/\bar{\alpha}, \bar{\alpha})$. Since the mean and the median coincide for the Student distribution, we have $\hat{\mu} = \bar{m} = 16.36$. For σ, the marginal posterior is $\pi(\sigma|y) \sim IG(\bar{\alpha}/2, \bar{\delta}/2)$. Using the 0.5 quantile of the inverse Gamma distribution yields the point estimate $\hat{\sigma} = 7.64$.

4.5 Credibility intervals

Another important concept in inference is that of estimation interval. The Bayesian intervals are known as credibility intervals and represent the counterparts of the frequentist confidence intervals.

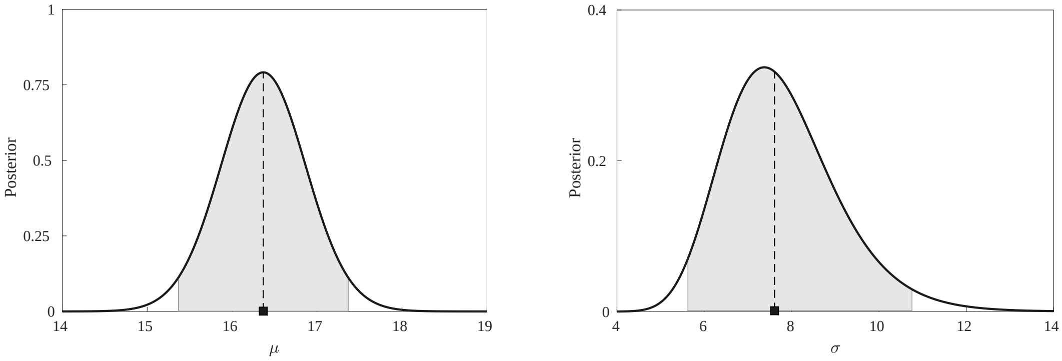
definition 4.5: let θ be some parameter; a **credibility interval** of level α is an interval of the form:
$\mathbb{P}(\theta_L \leq \theta \leq \theta_U|y) = 1-\alpha$
where θ_L and θ_U respectively denote the lower and upper bounds of the interval.

In other words, the Bayesian credibility interval is an interval that contains $(1-\alpha)\%$ of the posterior distribution of θ. Even though the credibility and confidence intervals may look similar, they differ fundamentally in conception. First, a confidence interval only integrates information from the data, while a Bayesian credibility interval also integrates the prior information. Second, and most importantly, the two methods consider the parameter θ differently. The frequentist approach treats θ as fixed and the confidence interval as random, hoping it contains the true parameter value with a probability $(1-\alpha)\%$. By constrast, the Bayesian credibility interval treats the interval boundaries as fixed and the parameter θ as random, the credibility region only delimiting a range that contains $(1-\alpha)\%$ of the posterior distribution $\pi(\theta|y)$.

In general, many different credibility intervals are possible for a given level α. One possibility consists in using the shortest possible interval, known as the highest posterior density interval. Finding this shortest interval may however prove computationally demanding. Often a simpler solution consists in building an equal-tail interval, that is, an interval that defines θ_L as the $\alpha/2$ quantile and θ_U as the $1-\alpha/2$ quantile of the posterior distribution $\pi(\theta|y)$. This choice is appealing when one uses the median (the 0.5 quantile) as a point estimate, for then it guarantees that the point estimate lies within the credibility interval.

example 4.1 (continued): we want to estimate credibility intervals for the parameters μ and σ in the stock return example, using the marginal posteriors developed in section 4.3. We use equal-tail intervals and set $\alpha = 0.05$ to obtain 95% credibility intervals. For μ, the marginal posterior is $\pi(\mu|y) \sim T(\bar{m}, \bar{\delta}\bar{v}/\bar{\alpha}, \bar{\alpha})$. We use the quantiles of the Student distribution to obtain $\mu_L = 15.25$ and $\mu_U = 17.48$. For σ, the marginal posterior is $\pi(\sigma|y) \sim IG(\bar{\alpha}/2, \bar{\delta}/2)$. Using the quantiles of the inverse Gamma, we obtain $\sigma_L = 5.62$ and $\sigma_U = 10.76$.

The marginal posterior distributions along with their point estimates and credibility intervals are depicted in Figure 4.2:



(a) Point estimate and interval for μ

(b) Point estimate and interval for σ

Figure 4.2: Point estimates and credibility intervals for μ and σ

4.6 The marginal likelihood

Often, we are interested in evaluating the overall goodness of fit of our model to the data. In this respect, the marginal likelihood $f(y)$ plays an important role in Bayesian analysis. Recall from definition 3.1 that the marginal likelihood represents the unconditional data density. In other words it provides a measure of the data likelihood regardless of the value of θ, and thus an assesment of the model in general.

The marginal likelihood $f(y)$ should not be confused with the likelihood function $f(y|\theta)$. There exists in fact a tight relation between the two concepts. From definitions 2.11 and 2.12, it follows that $f(y) = \int f(y,\theta)d\theta = \int f(y,\theta)/\pi(\theta) \times \pi(\theta)d\theta = \int f(y|\theta)\pi(\theta)d\theta$. In other words, the marginal likelihood represents the expectation of the likelihood function $f(y|\theta)$ over the prior distribution $\pi(\theta)$, that is, the average fit of the data over the prior belief.

definition 4.6: let $f(y|\theta)$ and $\pi(\theta)$ respectively denote the likelihood function and the prior distribution for some parameter θ. The **marginal likelihood** , denoted by $f(y)$, is given by:
$f(y) = \int f(y|\theta)\pi(\theta)d\theta$

Unlike Bayes rule where it is possible to work with kernels only, the marginal likelihood requires the inclusion of the normalization constants. Calculating the marginal likelihood can be tricky and sometimes impossible, but for simple models it can be obtained from direct application of definition 4.6.

example 4.1 (continued): we want to calculate the marginal likelihood for the stock return example, using the hierarchical prior developed in section 4.2. Applying definition 4.6, we obtain:

$$f(y) = \int\int f(y|\mu,\sigma)\ \pi(\mu|\sigma)\ \pi(\sigma)\ d\mu d\sigma \tag{1.4.14}$$

Using (1.3.19), (1.4.3) and (1.4.1), the expression becomes:

$$\begin{aligned} f(y) &= \int\int (2\pi\sigma)^{-n/2}\ \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma}\right) \\ &\times (2\pi v\sigma)^{-1/2}\exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v\sigma}\right)\ \times\ \frac{\delta/2^{\alpha/2}}{\Gamma(\alpha/2)}\sigma^{-\alpha/2-1}\exp\left(-\frac{\delta}{2\sigma}\right)d\mu d\sigma \end{aligned} \tag{1.4.15}$$

Note that unlike the posterior distribution, the marginal likelihood requires inclusion of the normalization constants. After some rearrangement and completing the squares, the expression becomes (book 2, p. 9):

$$\begin{aligned} f(y) &= \pi^{-n/2}\ (1+vn)^{-1/2}\ \frac{\delta^{\alpha/2}}{\bar{\delta}^{\bar{\alpha}/2}}\ \frac{\Gamma(\bar{\alpha}/2)}{\Gamma(\alpha/2)} \\ &\times\int\int (2\pi\bar{v}\sigma)^{-1/2}\exp\left(-\frac{1}{2}\frac{(\mu-\bar{m})^2}{\sigma\bar{v}}\right)\times\frac{\bar{\delta}/2^{\bar{\alpha}/2}}{\Gamma(\bar{\alpha}/2)}\sigma^{-\bar{\alpha}/2-1}\exp\left(-\frac{\bar{\delta}}{2\sigma}\right)d\mu d\sigma \end{aligned} \tag{1.4.16}$$

$\bar{m}, \bar{v}, \bar{\alpha}$ and $\bar{\delta}$ are defined as in (1.4.6). The expression may look messy, but the terms in the integral respectively represent the density function of a normal distribution and an inverse Gamma distribution. Therefore they both integrate to unity, only leaving the simple expression:

$$f(y) = \pi^{-n/2}\ (1+vn)^{-1/2}\ \frac{\delta^{\alpha/2}}{\bar{\delta}^{\bar{\alpha}/2}}\ \frac{\Gamma(\bar{\alpha}/2)}{\Gamma(\alpha/2)} \tag{1.4.17}$$

It is customary to reformulate the marginal likelihood in base 10 logarithm as $m(y) = \log_{10}(f(y))$. Let us now calculate the marginal likelihood for the stock return example. Given (1.4.17) and the values used in section 4.3, we obtain $m(y) = -26.07$. There is no direct interpretation for this value, but in the incoming section we will see how the marginal likelihood can be used to run model comparison and hypothesis testing.

4.7 Hypothesis testing and model comparison

In statistics, we are often interested in evaluating two competing hypotheses in light of the data, and then take a decision about which to accept. In a Bayesian context, hypothesis testing is straightforward. Given two competing hypotheses and some observed data, we first specify separate prior distributions to quantitatively describe each hypothesis. Combining the likelihood function for the data with each of the prior distributions, we obtain hypothesis-specific models. The overall goodness of fit of the model with the data under each hypothesis is then established from the marginal likelihood. Bayesian hypothesis testing thus amounts to finding the model best supported by the data through the marginal likelihood criterion.

It is worth noting that the procedure is general and is not restricted to hypothesis testing. It can be used for model comparison in general, even if the models are characterized by different parameters, priors, variables, and so on. Concretly, assume we want to compare two models M_1 and M_2, possibly corresponding to two competing hypotheses. For $i = 1,2$, we want to establish the probability that M_i is the correct model, given the data. This is obtained from the conditional probability $\mathbb{P}(M_i|y)$. Applying Bayes rule 3.1, it can be expressed as:

$$\mathbb{P}(M_i|y) = \frac{f(y|M_i)\ \mathbb{P}(M_i)}{f(y)} \tag{1.4.18}$$

$f(y|M_i)$ is the likelihood function under model M_i, and $\mathbb{P}(M_i)$ represents the prior belief that model M_i is indeed the correct model. $f(y)$ is the overall marginal likelihood, that is, the data density regardless of the model chosen. After basic manipulations, equation (1.4.18) reformulates as (book 2, p. 10):

$$\mathbb{P}(M_i|y) = \frac{\mathbb{P}(M_i)\ f_i(y)}{f(y)} \qquad f_i(y) \equiv \int f(y|M_i,\theta_i)\ \pi(\theta|M_i)d\theta_i \tag{1.4.19}$$

The numerator is constituted of two terms. The first term is the prior belief $\mathbb{P}(M_i)$ that model M_i is the correct one. The second term $f_i(y)$ can be recognised from definition 4.6 as the marginal likelihood for model M_i. To compare the two models, we simply take the ratio of the posterior probabilities.

definition 4.7: the **posterior odds** between models M_1 and M_2 is given by:

$$K = \frac{\mathbb{P}(M_1|y)}{\mathbb{P}(M_2|y)} = \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)}\frac{f_1(y)}{f_2(y)}$$

The ratio $\frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)}$ is known as the **prior odds**, while the ratio $\frac{f_1(y)}{f_2(y)}$ is the **Bayes factor**.

We see that model comparison reduces to a simple formula. First, it takes into account the prior odds, which reflects our prior belief about which model M_i is correct. In practice, the uninformative choice $\mathbb{P}(M_1) = \mathbb{P}(M_2) = 0.5$ is often made, in which case the posterior odds reduces to the Bayes factor. A larger value of the Bayes factor then indicates the the data is more supportive of model M_1, while values close to 1 indicate that both models are supported equally well. To decide on whether evidence is conclusive, Jeffreys (1961) propose to consider the value $\log_{10}(K) = m_1(y) - m_2(y)$, with $m_i(y) = \log_{10}(f_i(y))$. He provides the following guidelines:

$\log_{10}(K)$	evidence strength
$\log_{10}(K) < 0$	negative evidence (supports M_2)
$0 \leq \log_{10}(K) < 1/2$	weak evidence for M_1
$1/2 \leq \log_{10}(K) < 1$	substantial evidence for M_1
$1 \leq \log_{10}(K) < 3/2$	strong evidence for M_1
$3/2 \leq \log_{10}(K) < 2$	very strong evidence for M_1
$\log_{10}(K) \geq 2$	decisive support for M_1

Table 4.1: Jeffrey's Guidelines

example 4.1 (continued): assume the investor has the same prior belief as in section 4.3: an average annual return of \$12.7 with a variance of 0.1. He might change his investment strategy if the return proves significantly higher, at a level of \$15 with a variance of 0.1. We thus test the two competing hypotheses by comparing the model M_1 with $m = 15$ and $v = 0.1$ and the model M_2 with $m = 12.7$ and $v = 0.1$. Using the uninformative choice $\mathbb{P}(M_1) = \mathbb{P}(M_2) = 0.5$, the test reduces to the Bayes factor. Using (1.4.17), we obtain $m_1(y) = -22.28$ and $m_2(y) = -26.07$ so the test value is $\log_{10}(K) = 3.78$. There is decisive support for M_1 and the investor decides to change his investment strategy.

4.8 Predictions

Given a statistical model, predicting new data values often represents a central concern. Concretely, for a given sample of data observations y, we want to predict some new unobserved data value $\hat{y}$. Because the context is Bayesian, this should translate into some conditional density $f(\hat{y}|y)$. Also, the prediction should take into account the underlying uncertainty about θ. This motivates the following formula:

$$f(\hat{y}|y) = \frac{f(\hat{y},y)}{f(y)} = \int \frac{f(\hat{y},y,\theta)}{f(y)} d\theta = \int \frac{f(\hat{y},y,\theta)}{f(y,\theta)} \frac{f(y,\theta)}{f(y)} d\theta = \int f(\hat{y}|y,\theta)\pi(\theta|y)d\theta \tag{1.4.20}$$

where use has been made of definitions 2.11 and 2.12. We can see that the conditional density takes a convenient form. It represents the expectation of the density function $f(\hat{y}|y,\theta)$ for the unobserved data $\hat{y}$ over the posterior distribution $\pi(\theta|y)$.

definition 4.8: let $\hat{y}$ be some new unobserved data value; the **posterior predictive distribution** $f(\hat{y}|y)$ is given by:

$$f(\hat{y}|y) = \int f(\hat{y}|y,\theta)\ \pi(\theta|y)\ d\theta$$

where $f(\hat{y}|y,\theta)$ denotes the likelihood function for the predicted value $\hat{y}$.

Forming a prediction then reduces to a basic application of definition 4.8. This yields a full posterior predictive distribution from which point estimates and credibility intervals can be obtained directly, using the methods developed in sections 4.4 and 4.5.

example 4.1 (continued): the investor now wants to predict the market return of the stock, using the hierarchical model developed in section 4.2. The prediction will integrate both the uncertainty about the average return μ and its volatility σ. From definition 4.8, the posterior predictive distribution obtains from:

$$f(\hat{y}|y) = \int\int f(\hat{y}|y,\mu,\sigma)\pi(\mu,\sigma|y)d\mu d\sigma \tag{1.4.21}$$

Given equation (1.3.18), the likelihood function $f(\hat{y}|y,\mu,\sigma)$ for the predicted value $\hat{y}$ is given by:

$$f(\hat{y}|y,\mu,\sigma) = (2\pi\sigma)^{-1/2}\exp\left(-\frac{1}{2}\frac{(\hat{y}-\mu)^2}{\sigma}\right) \tag{1.4.22}$$

Combining with the posterior $\pi(\mu,\sigma|y)$ given by (1.4.4) and relegating to the normalization constant any term not involving $\hat{y},\mu$ or σ yields:

$$\begin{aligned} f(\hat{y}|y) \propto \int\int\ & \sigma^{-1/2}\exp\left(-\frac{1}{2}\frac{(\hat{y}-\mu)^2}{\sigma}\right) \times \sigma^{-n/2}\ \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma}\right) \\ & \times \sigma^{-1/2}\exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v\sigma}\right) \times \sigma^{-\alpha/2-1}\exp\left(-\frac{\delta}{2\sigma}\right) d\mu d\sigma \end{aligned} \tag{1.4.23}$$

This is not a pretty formula, but after some manipulations (book 2, p. 10) it can be expressed as:

$$f(\hat{y}|y) \propto \left(1+\frac{1}{\bar{\alpha}}\frac{(\hat{y}-\bar{m})^2}{\bar{\delta}(1+\bar{v})/\bar{\alpha}}\right)^{-(\bar{\alpha}+1)/2} \tag{1.4.24}$$

where $\bar{m},\bar{v},\bar{\alpha}$ and $\bar{\delta}$ are defined as in (1.4.6). This is recognised as the kernel of a Student distribution with location $\bar{m}$, scale $\bar{\delta}(1+\bar{v})/\bar{\alpha}$ and degrees of freedom $\bar{\alpha}$: $f(\hat{y}|y) \sim T(\bar{m},\bar{\delta}(1+\bar{v})/\bar{\alpha},\bar{\alpha})$.

Using the numerical values obtained in section 4.3, we obtain a posterior predictive density with location $\bar{m} = 16.36$, scale $\bar{\delta}(1+\bar{v})/\bar{\alpha} = 9.46$ and degrees of freedom $\bar{\alpha} = 47$. This yields a median point estimate of 16.36, and a 95% credibility interval of $\hat{y}_L = 10.18$ and $\hat{y}_U = 22.56$. The distribution, along with its point estimate and credibility interval is depicted in Figure 4.3:

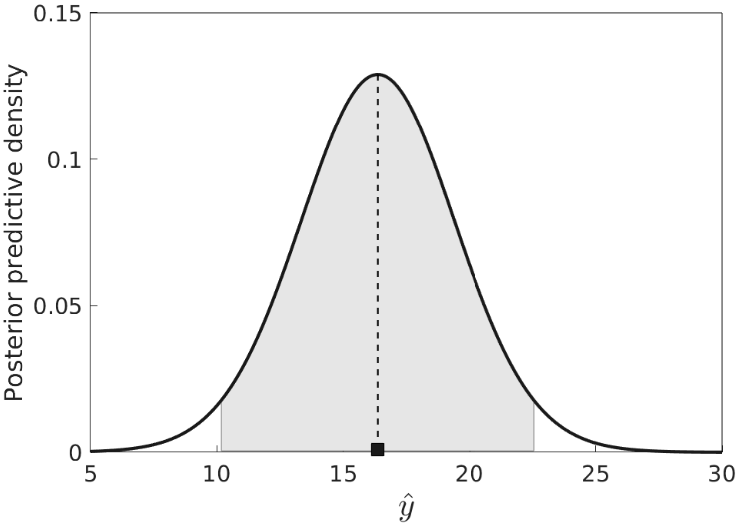


Figure 4.3: Posterior predictive distribution for the stock return example

From the distribution, the investor predicts a return comprised with 95% probability between 10.18 and 22.56, with a median forecast at 16.36.

CHAPTER 5

Properties of Bayesian estimates

This final introductory chapter focuses on the properties of posterior distributions. It provides further insights on their behaviours, and considers specifically the impact of the sample size and the specification of the prior distribution.

5.1 Posterior distribution as a compromise between prior and likelihood

The posterior distribution involves the combination of the prior distribution with the likelihood function. It is therefore natural to expect that since it contains the information from both sources, it will appear as a compromise between them. This is in fact true, and contitutes a general feature of Bayesian inference. The three applied examples developed in chapter 3 made this point apparent, especially when looking at Figures 3.2, 3.3 and 3.4 which all show the posterior between the prior and the likelihood function. We now make this point formal by looking at the example algebra.

example 5.1: consider again the coin flip example developed in section 3.2. The posterior distribution is $\pi(p|y) \sim Beta(\bar{\alpha},\bar{\beta})$, with $\bar{\alpha} = \alpha + m$ and $\bar{\beta} = \beta + n - m$. Denoting the posterior mean by $\mathbb{E}(p|y)$, the prior mean by $\mathbb{E}(p)$ and the maximum likelihood estimate by $\hat{p}$, it can be shown that (book 2, p. 15):

$$\mathbb{E}(p|y) = \gamma\ \mathbb{E}(p) + (1-\gamma)\ \hat{p} \qquad \text{with} \qquad \gamma = \frac{\alpha+\beta}{\alpha+\beta+n} \tag{1.5.1}$$

In other words, the posterior mean is a weighted average between the prior mean and the maximum likelihood estimate, the weight being defined by the hyperparameters α and β and the sample size n.

example 5.2: consider again the car sale example developed in section 3.3. The posterior distribution is $\pi(\lambda|y) \sim G(\bar{a},\bar{b})$, with $\bar{a} = a + \sum_{i=1}^{n} y_i$ and $\bar{b} = \frac{b}{bn+1}$. It can then be shown that (book 2, p. 15):

$$\mathbb{E}(\lambda|y) = \gamma\ \mathbb{E}(\lambda) + (1-\gamma)\ \hat{\lambda} \qquad \text{with} \qquad \gamma = \frac{1}{bn+1} \tag{1.5.2}$$

The posterior mean is a weighted average between the prior mean and the maximum likelihood estimate, the weight being defined by the hyperparameter b and the sample size n.

example 5.3: consider again the stock return example developed in section 3.4, assuming that μ is the only parameter to estimate. The posterior distribution is $\pi(\mu|y) \sim N(\bar{m},\bar{v})$, with $\bar{v} = \left(\frac{n}{\sigma} + \frac{1}{v}\right)^{-1}$ and $\bar{m} = \bar{v}\left(\frac{1}{\sigma}\sum_{i=1}^{n} y_i + \frac{m}{v}\right)$. It can then be shown that (book 2, p. 16):

$$\mathbb{E}(\mu|y) = \gamma\ \mathbb{E}(\mu) + (1-\gamma)\ \hat{\mu} \qquad \text{with} \qquad \gamma = \frac{\sigma}{vn+\sigma} \tag{1.5.3}$$

The posterior mean is a weighted average between the prior mean and the maximum likelihood estimate, the weight being defined by the variance σ, the hyperparameter v and the sample size n.

These results are summarised in Table 5.1, along with the posterior variances of the parameters.

Example	parameter	prior mean	MLE	posterior mean	weight γ	posterior variance
coin flip	p	$\frac{\alpha}{\alpha+\beta}$	$\frac{m}{n}$	$\gamma\mathbb{E}(p)+(1-\gamma)\hat{p}$	$\frac{\alpha+\beta}{\alpha+\beta+n}$	$\frac{(\alpha+m)(\beta+n-m)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$
car sales	λ	ab	$\frac{1}{n}\sum_{i=1}^{n} y_i$	$\gamma\mathbb{E}(\lambda)+(1-\gamma)\hat{\lambda}$	$\frac{1}{bn+1}$	$\frac{(a+\sum_{i=1}^{n} y_i)b^2}{(bn+1)^2}$
stock return	μ	m	$\frac{1}{n}\sum_{i=1}^{n} y_i$	$\gamma\mathbb{E}(\mu)+(1-\gamma)\hat{\mu}$	$\frac{\sigma}{vn+\sigma}$	$\frac{\sigma}{n+\sigma/v}$

Table 5.1: Posteriors as weigthed average of prior mean and maximum likelihood estimate

In our three examples it was possible to represent the posterior mean $\mathbb{E}(\theta|y)$ as a weighted average $\mathbb{E}(\theta|y) = \gamma\ \mathbb{E}(\theta) + (1-\gamma)\ \hat{\theta}$ of the maximum likelihood estimate $\hat{\theta}$ and the prior mean $\mathbb{E}(\theta)$. Is it always possible to do so? Diaconis and Ylvisaker (1979) show that the answer is yes for conjugate priors belonging to the family of exponential distributions. This family comprises many common distributions including the normal, Beta, and Gamma distributions.

For other priors, the posterior mean may possibly not be expressed in that form. Even in this case, the posterior distribution remains a compromise between the prior information and the data, with its center somewhere in-between. How much weight is attributed to each component then depends on the sample size and the prior tightness, as developed in the incoming sections.

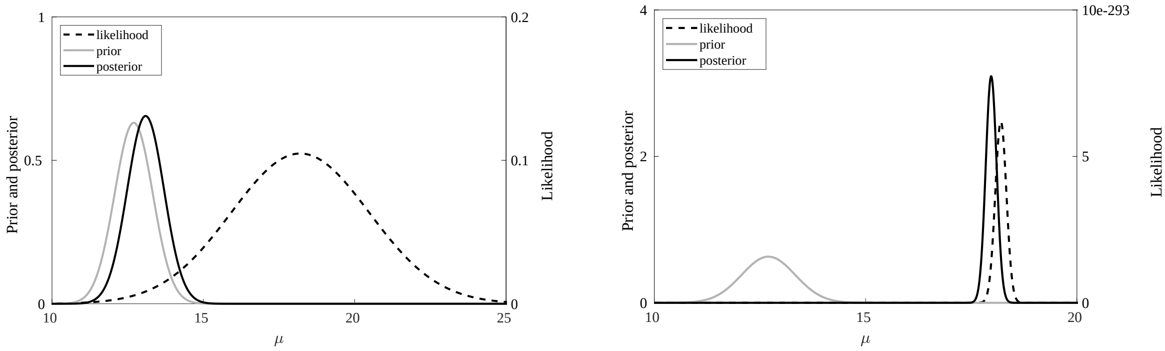
5.2 Large VS. small samples

We now consider the impact of the sample size on the posterior distribution. Intuitively, a large sample means a large amount of data information relative to that contained in the prior. Following, we expect the posterior to reflect more the likelihood function than the prior distribution. This is indeed correct, and represents a fundamental feature of Bayesian estimates.

Consider the weight column of Table 5.1. It is easily seen that for all three examples the weight γ diminishes as n increases, pushing the posterior mean $\mathbb{E}(\theta|y)$ away from the prior mean $\mathbb{E}(\theta)$ and towards the maximum likelihood estimate $\hat{\theta}$. In the limit case where $n \to \infty$, we have $\gamma \to 0$ and the posterior mean coincides with the maximum likelihood estimate. Conversely, when $n \to 0$ we see that $\gamma \to 1$ and $\mathbb{E}(\theta|y) \to \mathbb{E}(\theta)$. This is because there is no data information at all so that all the weight goes to the prior.

Interestingly enough, the sample size n also impacts the posterior variance. Looking at the final column of Table 5.1, we see that as n increases the posterior variance diminishes for the three examples. As $n \to \infty$ the posterior variance tends to 0 and the posterior distribution collapses to a single mass point at the maximum likelihood estimate $\hat{\theta}$. On the other hand, when $n \to 0$ the posterior variance converges to the prior variance. In fact, in this case, the posterior distribution as a whole converges to the prior distribution. This is again due to the absence of data information which leaves only the prior to carry the estimation.

These properties are illustrated in Figure 5.1 which plots the likelihood, prior and posterior for the stock return example with sample size $n = 1$ on the left and $n = 300$ on the right.



(a) Small sample ($n = 1$)

(b) Large sample ($n = 300$)

Figure 5.1: Likelihood, prior and posterior with small and large samples

On the left panel (small sample) the posterior distribution matches the prior distribution almost perfectly. The likelihood function is widely spread, reflecting imprecise information about the parameter. On the right panel (large sample) the posterior gets much closer to the likelihood function, as the latter now provides most of the available information on the parameter. It becomes also tighter, reflecting improved accuracy through the larger number of observations.

To summarize:

When n is small:

- the likelihood function plays a negligible role, and most of the weight is given to the prior distribution.
- the posterior mean and variance converge to their prior counterparts.
- the posterior distribution is identical to the prior.

When n is large:

- the prior becomes marginal, and most of the weight goes to the likelihood function.
- the posterior mean converges to the maximum likelihood estimator.
- the posterior variance tends to 0, and the posterior as a whole becomes a degenerate distribution with a mass point at the maximum likelihood estimator.

5.3 Informative VS. uninformative priors

The prior distribution reflects our subjective belief about the parameter θ. We may be confident in this prior belief, in which case we want the prior distribution $\pi(\theta)$ to be granted much weight. On the contrary we may have only vague knowledge about θ, in which case we want to put little weight to the prior and leave most of the decision to the data, i.e. the likelihood function $f(y|\theta)$.

definition 5.1: an **uninformative prior** or **diffuse prior** is a prior distribution $\pi(\theta)$ that reflects vague or nonexistent knowledge about the parameter θ. The distribution contains no prior information and leaves the burden of estimation entirely to the data.

The informativeness of a prior distribution $\pi(\theta)$ is directly related to the prior variance $Var(\theta)$. A tight prior distribution means that we are very confident in our prior information, so that much weight is attributed to the prior. In the limit case where $Var(\theta) \to 0$, the posterior $\pi(\theta|y)$ converges to the prior $\pi(\theta)$.

By contrast, a loose prior distribution implies vague or imprecise knowledge of θ and translates into a large prior variance. In this case, the data will represent the bulk of the information and the posterior will attribute all the weight to the likelihood function. In the limit case where $Var(\theta) \to \infty$, the prior becomes uninformative and the posterior distribution $\pi(\theta|y)$ converges to the likelihood function $f(y|\theta)$.

An extreme way to generate uninformative priors is to use an improper prior, a prior that is not integrable and exhibits infinite variance.

definition 5.2: an **improper prior** is a prior distribution $\pi(\theta)$ whose integral is infinity. By contrast, a prior distribution whose integral is unity is called a **proper prior**.

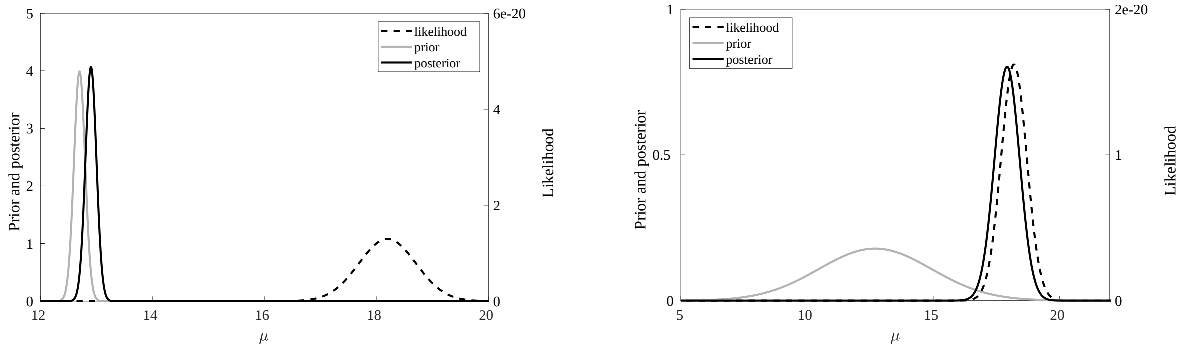
For instance, to specify a prior distribution on some parameter θ taking real values, we may use the improper prior $\pi(\theta) \propto 1$. This defines a uniform distribution over the interval $[-\infty, +\infty]$. The distribution integrates to infinity and connot be normalised to one. Improper priors will typically yield proper posteriors, which makes them appealing to reflect agnostic prior beliefs. However, they prevent the calculation of the marginal likelihood which requires the normalization constants. In this respect, it is preferable to specify a proper prior (even weakly informative) whenever possible.

To illustrate these properties, consider again the weight column of Table 5.1. For the coin flip example, a diffuse Beta prior for p can be obtained by setting $\alpha \to 0$ and $\beta \to 0$. In this case the weight γ tends to 0 and all the weight get to the maximum likelihood estimate $\hat{p}$. On the other hand, setting $\alpha \to \infty$ and $\beta \to \infty$ results in a very tight prior and in this case it is easily seen that γ tends to 1, attributing all the weight to the prior distribution.

For the car sales example, a diffuse Gamma prior can be obtained by setting $b \to \infty$, in which case it is easily seen that γ tends to 0. Conversely, a tight prior obtains by setting $b \to 0$, resulting in the weight γ tending to 1.

For the stock return example, the prior variance is just v. An uninformative prior can then be obtained by setting $v \to \infty$, and then γ tends to 0. Conversely, with $v \to 0$ the prior gets informative and γ tends to 1, putting all the weight on the prior.

These properties are illustrated in Figure 5.2 which plots the likelihood, prior and posterior for the stock return example with prior variance $v = 0.01$ on the left and $v = 5$ on the right.



(a) Small prior variance ($v = 0.01$)

(b) Large prior variance ($v = 5$)

Figure 5.2: Likelihood, prior and posterior with small and large prior variance

On the left panel (small prior variance), the tight prior reflects a strong confidence in the prior belief. Following, all the weight goes to the prior and the posterior distribution matches it almost perfectly. On the right panel (large prior variance), the prior is seen to be widely spread, reflecting the lack of accurate information. This pushes the posterior towards the likelihood function, the burden of estimation now being exclusively on the data.

To summarize:

When $Var(\theta)$ is small:

- the likelihood function plays a negligible role, and most of the weight is given to the prior distribution.
- the posterior mean and variance converge to their prior counterparts.
- the posterior distribution is identical to the prior.

When $Var(\theta)$ is large:

- the prior becomes marginal, and most of the weight goes to the likelihood function.
- the posterior mean converges to the maximum likelihood estimator.

PART II

Simulation methods

CHAPTER 6

The Gibbs sampling algorithm

This chapter and the next one introduce the simulation methods that constitute the workhorse of modern Bayesian econometrics. This chapter focuses on the Gibbs sampling algorithm, the simplest approach whenever simulation methods are needed. Chapter 7 then discusses the Metropolis-Hastings algorithm, a more general but also more computationally intensive alternative. The two chapters adopt a cookbook approach: the methods are introduced without developing the underlying mathematical theory. The algebra behind the algorithms is introduced only in chapter 8, and the readers uninterested in mathematical details may safely skip this part.

6.1 Gibbs sampling: motivation

Consider again the stock return example introduced in chapter 3: an investor wants to predict the return of a given stock on the NYSE. The return is modelled as a normal distribution with mean μ and variance σ. Section 3.4 introduced the simplest version of the problem, assuming that only μ was unknown. In section 4.2 the problem was made more realistic by assuming that both μ and σ were unknown, and it was solved using a hieraractical prior. However, the hierarchical prior is undesirable because it relies on the strong assumption that the prior variance of μ is proportional to σ.

Ideally, μ and σ must be modelled as independent parameters, as was done in section 4.1. The priors for μ and σ are respectively given by $\pi(\mu) \sim N(m,v)$ (equation (1.3.22)) and $\pi(\sigma) \sim IG(\alpha/2, \delta/2)$ (equation (1.4.1)). Combined with the likelihood function (1.3.19) and applying Bayes rule, the joint posterior is given by equation (1.4.2), repeated here for convenience:

$$\pi(\mu,\sigma|y) \propto \sigma^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma}\right) \times \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v}\right) \times \sigma^{-\alpha/2-1}\exp\left(-\frac{\delta}{2\sigma}\right) \tag{2.6.1}$$

Following definition 4.3, we would like to obtain the marginal posteriors from $\pi(\mu|y,\sigma) = \int \pi(\mu,\sigma|y)d\sigma$ and $\pi(\sigma|y,\mu) = \int \pi(\mu,\sigma|y)d\mu$. However, as already stated at the end of section 4.3, it is not possible to calculate these integrals as the μ and σ terms are too interwoven to permit marginalization.

This example shows the limits of traditional Bayesian methods: even though the model is simple and involves only two parameters, it does not have closed forms solutions for its posterior distributions. Because of such difficulties, Bayesian econometrics up to the 1970's was essentially restricted to trivial conjugate models with easily evaluable posteriors. In the 1980's, as more sophisticated methods became available under the frequentist approach, interest in Bayesian methods gradually declined. This changed in the 1990's with the dramatic rise in computing power. Simulation methods that were unimaginable in the 1970's became trivially accessible with modern computers. This eventually led to the developement of the so-called Markov Chain Monte Carlo methods (often abbreviated as MCMC methods), and in particular the Gibbs sampling algorithm.

6.2 Gibbs sampling: the algorithm

Whenever analytical solutions are unavailable, it may still be possible to evaluate the marginal posteriors numerically. By this we mean that it is possible to sample values from the marginal posterior distributions even though their analytical form is unknown. By sampling sufficiently many values, one obtains an empirical distribution that approximates the real distribution and can be used to obtain empirical point estimates, credibility intervals, and so on.

The Gibbs sampling algorithm represents the simplest approach to simulation methods. It is available whenever the conditional posteriors are known distributions from which it is possible to sample values. Consider a model with n parameters $\theta = \{\theta_1, \cdots, \theta_n\}$, joint posterior distribution $\pi(\theta_1, \cdots, \theta_n | y)$, and conditional posteriors $\pi(\theta_1 | y, \theta_2, \cdots, \theta_n)$, $\cdots$, $\pi(\theta_n | y, \theta_1 \cdots, \theta_{n-1})$ (we will see soon how to derive these conditional posteriors). Assume the conditional posteriors are known distributions so that we can easily sample values from them. The Gibbs sampling algorithm then consists in:

algorithm 6.1: Gibbs sampling algorithm

1. set any initial values $\theta_1^{(0)}, \cdots, \theta_n^{(0)}$ for the n parameters (these initial values are unimportant for the rest of the algorithm).

2. at the first iteration, draw:

$$\begin{gathered} \theta_1^{(1)} \text{ from } \pi(\theta_1 | y, \theta_2^{(0)}, \cdots, \theta_n^{(0)}) \\ \theta_2^{(1)} \text{ from } \pi(\theta_2 | y, \theta_1^{(1)}, \cdots, \theta_n^{(0)}) \\ \vdots \\ \theta_n^{(1)} \text{ from } \pi(\theta_n | y, \theta_1^{(1)}, \cdots, \theta_{n-1}^{(1)}) \end{gathered}$$

3. at iteration j, draw:

$$\begin{gathered} \theta_1^{(j)} \text{ from } \pi(\theta_1 | y, \theta_2^{(j-1)}, \cdots, \theta_n^{(j-1)}) \\ \theta_2^{(j)} \text{ from } \pi(\theta_2 | y, \theta_1^{(j)}, \cdots, \theta_n^{(j-1)}) \\ \vdots \\ \theta_n^{(j)} \text{ from } \pi(\theta_n | y, \theta_1^{(j)}, \cdots, \theta_{n-1}^{(j)}) \end{gathered}$$

4. repeat until the desired number of iterations is realised.

The principle behind the algorithm is simple: draw sequentially the parameters $\theta_1, \cdots, \theta_n$ from their conditional posteriors distributions $\pi(\theta_1 | y, \theta_2, \cdots, \theta_n)$, $\cdots$, $\pi(\theta_n | y, \theta_1 \cdots, \theta_{n-1})$, and repeat the process a large number of times. After a certain number of iterations, the algorithm converges to the target distributions which are the marginal posterior distributions $\pi(\theta_1 | y)$, $\cdots$, $\pi(\theta_n | y)$.

In technical terms, we say that after a sufficient number of iterations known as the **transient sample** or **burn-in sample**, the algorithm converges to the **invariant distribution** of the Markov Chain, which is just the set of marginal posteriors $\pi(\theta_1 | y)$, $\cdots$, $\pi(\theta_n | y)$. The order in which the parameters $\theta_1, \cdots, \theta_n$ are drawn within each iteration is unimportant, which may sometimes prove convient. Note Also that the existence of a transient sample implies that the initial draws are not sampled from the invariant distribution and must then be discarded.

These remarkable convergence properties constitute the core of modern numerical methods applied to Bayesian analysis. Their only requirements are knowledge of the conditional posterior distributions for the model, and sufficient computer speed to accomplish the steps. It now remains to discuss how the conditional posteriors can be obtained. Using definition 2.12, and denoting by $\theta_{j \neq i}$ the set of all parameters except θ_i, it follows directly that:

$$\pi(\theta_i|y, \theta_{j\neq i}) = \frac{\pi(y, \theta_i, \theta_{j\neq i})}{\pi(y, \theta_{j\neq i})} = \frac{\pi(y, \theta)}{\pi(y, \theta_{j\neq i})} = \frac{\pi(y, \theta)}{f(y)} \frac{f(y)}{\pi(y, \theta_{j\neq i})} = \frac{\pi(\theta|y)}{\pi(\theta_{j\neq i}|y)} \propto \pi(\theta|y) \tag{2.6.2}$$

The final step obtains by noting that the joint posterior $\pi(\theta_{j\neq i}|y)$ does not involve θ_i and can thus be relegated to the normalization constant. What equation (2.6.2) shows is that the conditional posterior $\pi(\theta_i|y, \theta_{j\neq i})$ is simply proportional to the joint posterior $\pi(\theta|y)$.

definition 6.1: let $\pi(\theta|y)$ denote the joint posterior for $\theta = \{\theta_1, \cdots, \theta_n\}$. The **conditional posterior distribution** $\pi(\theta_i|y, \theta_{j\neq i})$ for θ_i obtains from:

$\pi(\theta_i|y, \theta_{j\neq i}) \propto \pi(\theta|y)$

In other words, to obtain $\pi(\theta_i|y, \theta_{j\neq i})$, one simply starts from the joint posterior $\pi(\theta|y)$ and relegate to the normalization constant any term not involving θ_i. If this yields a known distribution, one can use the gibbs sampling algorithm to sample directly from $\pi(\theta_i|y)$.

6.3 Gibbs sampling: an example

We now illustrate the use of the Gibbs sampling algorithm with the stock return example. Consider the joint posterior (2.6.1). To use the Gibbs sampling algorithm, we need the conditional posteriors $\pi(\mu|y, \sigma)$ and $\pi(\sigma|y, \mu)$.

Consider the conditional posterior $\pi(\mu|y, \sigma)$. Using definition 6.1, start from the joint posterior $\pi(\mu, \sigma|y)$ given by (2.6.1) and relegate to the normalization constant any term not involving μ. Doing so yields:

$$\pi(\mu|y, \sigma) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma}\right) \times \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v}\right) \tag{2.6.3}$$

This expression is similar to (1.3.23). Following the same approach and completing the squares eventually yields:

$$\pi(\mu|y, \sigma) \propto \exp\left(-\frac{1}{2}\frac{(\mu-\bar{m})^2}{\bar{v}}\right) \tag{2.6.4}$$

with:

$$\bar{v} = \left(\frac{n}{\sigma} + \frac{1}{v}\right)^{-1} \qquad \bar{m} = \bar{v}\left(\frac{1}{\sigma}\sum_{i=1}^{n} y_i + \frac{m}{v}\right) \tag{2.6.5}$$

This is the kernel of a normal distribution with mean $\bar{m}$ and variance $\bar{v}$: $\pi(\mu|y, \sigma) \sim N(\bar{m}, \bar{v})$.

Consider then the conditional posterior $\pi(\sigma|y, \mu)$. Start from the joint posterior $\pi(\mu, \sigma|y)$ given by (2.6.1) and relegate to the normalization constant any term not involving σ to obtain:

$$\pi(\sigma|y, \mu) \propto \sigma^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma}\right) \times \sigma^{-\alpha/2-1} \exp\left(-\frac{\delta}{2\sigma}\right) \tag{2.6.6}$$

Rearranging the terms in (2.6.6) directly yields:

$$\pi(\sigma|y,\mu) \propto \sigma^{-\bar{\alpha}/2-1}\, \exp\left(-\frac{\bar{\delta}}{2\sigma}\right) \tag{2.6.7}$$

with:

$$\bar{\alpha} = \alpha + n \qquad \bar{\delta} = \delta + \sum_{i=1}^{n}(y_i - \mu)^2 \tag{2.6.8}$$

This is the kernel of an inverse Gamma distribution with shape $\bar{\alpha}/2$ and scale $\bar{\delta}/2$: $\pi(\sigma|y,\mu) \sim IG(\bar{\alpha}/2, \bar{\delta}/2)$.

From direct application of algorithm 6.1, the Gibbs sampler for the model obtains as:

algorithm 6.2: Gibbs sampling algorithm for the stock return model

1. set initial values $\mu^{(0)}$ and $\sigma^{(0)}$. We use the sample estimates $\mu^{(0)} = \hat{\mu}$ and $\sigma^{(0)} = \hat{\sigma}$.
2. at iteration j, draw:

   $\mu^{(j)}$ from $\pi(\mu|y,\sigma^{(j-1)}) \sim N(\bar{m},\bar{v})$ with:

   $$\bar{v} = \left(\frac{n}{\sigma^{(j-1)}} + \frac{1}{v}\right)^{-1} \qquad \bar{m} = \bar{v}\left(\frac{1}{\sigma^{(j-1)}}\sum_{i=1}^{n} y_i + \frac{m}{v}\right)$$

3. at iteration j, draw:

   $\sigma^{(j)}$ from $\pi(\sigma|y,\mu^{(j)}) \sim IG(\bar{\alpha}/2, \bar{\delta}/2)$ with:

   $$\bar{\alpha} = \alpha + n \qquad \bar{\delta} = \delta + \sum_{i=1}^{n}(y_i - \mu^{(j)})^2$$

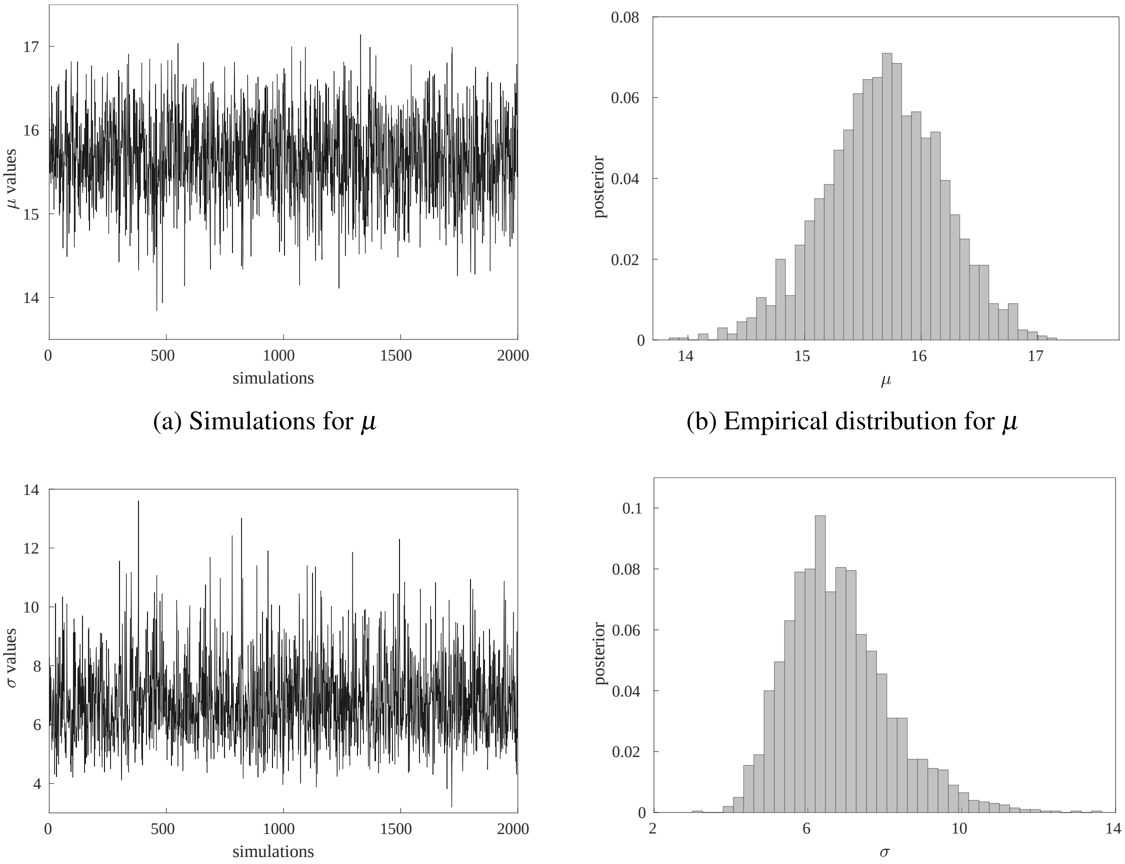
4. repeat to obtain 1000 iterations as burn-in sample and 2000 additional iterations for simulated values.

The resulting simulated values along with the associated empirical distributions are displayed in Figure 6.1. The left panels show the simulations obtained for the Gibbs sampler (after discarding the burn-in fraction), while the right panels show the resulting empirical distributions. These empirical distributions look close to the ones obtained analytically with the hierarchical prior (compare for instance with Figure 4.2).

Figure 6.1 also highlights the cost from using the Gibbs sampling approach: clearly, the empirical distributions are only approximate and don't exhibit the same degree of accuracy as their analytical counterparts. One reason for that here is the small number of simulations: with only 1000 burn-in iterations and 2000 sample iterations the distribution can only be rough. By increasing both values a more accurate distribution could be obtained, at the cost of increased computational time.

In general, there is no objective rule to determine how many burn-in and sample iterations should be used. More burn-in iterations improve convergence towards the invariant distribution, and more sample iterations produce a finer empirical distribution. On the other hand, depending on the model, the computational cost may become prohibitive. For most simple econometrics models, 1000 burn-in and 2000 sample iterations is typically enough, but more complex models may require many more iterations to obtain a reasonably fine empirical distribution, using dozens of thousands iterations as burn-in and sample.

Once the empirical distributions are obtained, they can be used for general purposes. For instance, we can use the empirical median to obtain point estimates and the 0.025 and 0.975 quantiles to compute the lower and upper bounds of a 95% credibility interval. Doing so, we find for μ a point estimate of 15.69 and a 95% credibility interval of $[14.65, 16.64]$. For σ, we obtain a point estimate of 6.65, and a 95% credibility interval of $[4.62, 9.91]$.



(a) Simulations for μ

(b) Empirical distribution for μ

(c) Simulations for σ

(d) Empirical distribution for σ

Figure 6.1: Gibbs sampling simulations and empirical distributions for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$

6.4 Posterior predictive distribution with Gibbs sampling

Recall from definition 4.8 that the posterior predictive distribution is given by:

$$f(\hat{y}|y) = \int f(\hat{y}|y,\theta)\ \pi(\theta|y)\ d\theta \tag{2.6.9}$$

Whenever one has to rely on simulation methods, this definition cannot be applied analytically because the exact form of the posterior $\pi(\theta|y)$ is unknown. Fortunately, it is straigthforward to adapt the definition to the simulation framework. Observing (2.6.9), we notice that the posterior predictive distribution writes as the product of the posterior $\pi(\theta|y)$, and the likelihood $f(\hat{y}|y,\theta)$ of future observations conditional on data y and parameters θ.

This suggests a direct simulation method to obtain draws from $f(\hat{y}|y)$. Suppose one can generate random draws for θ from the posterior $\pi(\theta|y)$, and then use this θ value to compute $\hat{y}$ from $f(\hat{y}|y,\theta)$. This produces a draw of $\hat{y}$ and θ from $f(\hat{y}|y,\theta)\ \pi(\theta|y)$. Marginalizing, which simply implies to discard the θ value then produces a draw from $\int f(\hat{y}|y,\theta)\ \pi(\theta|y)\ d\theta$, i.e. from $f(\hat{y}|y)$.

It is trivially simple to generate draws from $\pi(\theta|y)$ because we can just recycle the values obtained from the Gibbs sampling algorithm. This way, a full Gibbs sampling algorithm for the posterior predictive distribution can be obtained as:

algorithm 6.3: Gibbs sampling algorithm for the posterior predictive distribution

1. at iteration j, draw $\theta_1^{(j)}$ from $\pi(\theta_1|y)$, $\theta_2^{(j)}$ from $\pi(\theta_2|y), \cdots$, and $\theta_n^{(j)}$ from $\pi(\theta_n|y)$. Simply recycle the values $\theta_1^{(j)}, \theta_2^{(j)}, \cdots, \theta_n^{(j)}$ obtained from the j^{th} iteration of the Gibbs sampling algorithm.
2. given $\theta^{(j)}$, draw $\hat{y}^{(j)}$ from $f(\hat{y}|y, \theta^{(j)})$.
3. marginalize, that is, discard $\theta^{(j)}$ and keep only $\hat{y}^{(j)}$.
4. repeat until the desired number of iterations is realised.

Running this algorithm, we obtain a sample of draws $\hat{y}^{(1)}, \hat{y}^{(2)}, \cdots$ which can be used to obtain an empirical distribution.

Consider for example the predictive distribution for the stock return example with Gibbs sampling. From (1.4.22), the likelihood function $f(\hat{y}|y, \mu, \sigma)$ for the predicted value $\hat{y}$ is given by:

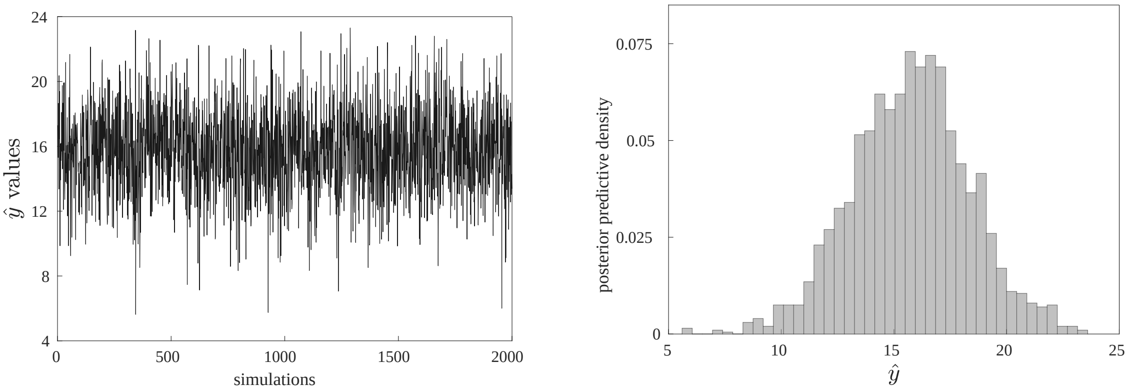
$$f(\hat{y}|y, \mu, \sigma) = (2\pi\sigma)^{-1/2} \exp\left(-\frac{1}{2}\frac{(\hat{y}-\mu)^2}{\sigma}\right) \tag{2.6.10}$$

In other words, the conditional distribution is normal with mean μ and variance σ: $f(\hat{y}|y, \mu, \sigma) \sim N(\mu, \sigma)$. This gives the following algorithm:

algorithm 6.4: Gibbs sampling algorithm for the posterior predictive distribution, stock return model

1. at iteration j, draw $\mu^{(j)}$ from $\pi(\mu|y)$ and $\sigma^{(j)}$ from $\pi(\sigma|y)$. Recycle the values $\mu^{(j)}$ and $\sigma^{(j)}$ obtained from the j^{th} iteration of the Gibbs sampling algorithm.
2. given $\mu^{(j)}$ and $\sigma^{(j)}$, draw $\hat{y}^{(j)}$ from $f(\hat{y}^{(j)}|y, \mu^{(j)}, \sigma^{(j)}) \sim N(\mu^{(j)}, \sigma^{(j)})$.
3. marginalize, that is, discard $\mu^{(j)}$ and $\sigma^{(j)}$, and keep only $\hat{y}^{(j)}$.
4. repeat until 2000 iterations are realised.

The simulated values and the associated empirical distributions are displayed in Figure 6.2.



(a) Simulations for $\hat{y}$ (b) Empirical distribution for $\hat{y}$

Figure 6.2: Gibbs sampling simulations and empirical distributions for $\hat{y}$

The empirical predictions are quite close to those obtained analytically with the hierarchical prior (compare with Figure 4.3). We can use the empirical distribution to obtain a point estimate of 15.81 and a 95% prediction interval of $[10.46, 21.04]$.

6.5 Marginal likelihood with Gibbs sampling

The marginal likelihood is normally calculated from definition 4.6 as $f(y) = \int f(y|\theta)\pi(\theta)d\theta$. However, the use of simulation methods implies that this quantity cannot be calculated, because the integral has no analytical solution. In this case, Chib (1995) proposes an alternative approach. First, rearranging Bayes rule 3.1 we obtain the identity:

$$f(y) = \frac{f(y|\theta)\pi(\theta)}{\pi(\theta|y)} \tag{2.6.11}$$

The numerator in (2.6.11) is known, since it is the product of the likelihood function $f(y|\theta)$ with the prior $\pi(\theta|y)$. The normalization constant of the denominator however is unknown so that $\pi(\theta|y)$ cannot be computed directly. The strategy consists in approximating the term from the values obtained from the Gibbs sampling algorithm.

Consider a two-parameter model with $\theta = \{\theta_1, \theta_2\}$. Using definition 2.12 of conditional densities, it follows that $\pi(\theta_1, \theta_2|y) = \pi(\theta_1|y, \theta_2)\pi(\theta_2|y)$. The first term is known: it is the conditional posterior $\pi(\theta_1|y, \theta_2)$, required for the Gibbs sampler. The second term is unknown, but can be reformulated as:

$$\pi(\theta_2|y) = \int \pi(\theta_2, \theta_1|y)d\theta_1 = \int \pi(\theta_2|\theta_1, y)\pi(\theta_1|y)d\theta_1 \tag{2.6.12}$$

The integral cannot be calculated analytically, but it can be approximated with the so-called importance sampling method: first sample J values $\theta_1^{(1)}, \cdots, \theta_1^{(J)}$ from $\pi(\theta_1|y)$, then compute the approximation:

$$\int \pi(\theta_2|\theta_1, y)\pi(\theta_1|y)d\theta_1 \approx \frac{1}{J}\sum_{j=1}^{J} \pi(\theta_2|\theta_1^{(j)}, y) \tag{2.6.13}$$

In practice, we use or course the J values generated by the Gibbs sampling algorithm. Substituting this formula back in (2.6.11), we find that the two-parameter marginal likelihood can be approximated by:

$$f(y) \approx \frac{f(y|\theta_1, \theta_2)\pi(\theta_1, \theta_2)}{\pi(\theta_1|y, \theta_2) \times \frac{1}{J}\sum_{j=1}^{J} \pi(\theta_2|\theta_1^{(j)}, y)} \tag{2.6.14}$$

The expression can be evaluated at any value of θ_1 and θ_2, but in general points of high density such as the median or the mode are chosen to optimize numerical accuracy. Denoting by $\theta^* = \{\theta_1^*, \theta_2^*\}$ the chosen high-density values, we eventually obtain:

$$f(y) \approx \frac{f(y|\theta_1^*, \theta_2^*)\pi(\theta_1^*, \theta_2^*)}{\pi(\theta_1^*|y, \theta_2^*) \times \frac{1}{J}\sum_{j=1}^{J} \pi(\theta_2^*|\theta_1^{(j)}, y)} \tag{2.6.15}$$

It is possible to switch θ_1^* and θ_2^* in (2.6.15), based on convenience. The methodology of Chib (1995) can be extended to models with more than two parameters, but the procedure gets considerably more complex and the computational cost may become prohibitive. In this case, simpler and more efficient alternatives may be prefered (see section 7.4).

We now apply the method to the stock return example. Given $\theta = \{\mu, \sigma\}$, (2.6.15) becomes:

$$f(y) \approx \frac{f(y|\mu^*, \sigma^*)\pi(\mu^*, \sigma^*)}{\pi(\sigma^*|y, \mu^*) \times \frac{1}{J}\sum_{j=1}^{J} \pi(\mu^*|y, \sigma^{(j)})} \tag{2.6.16}$$

To evaluate (2.6.16), we use the likelihood function $f(y|\mu,\sigma)$ given by (1.3.19), the priors $\pi(\mu)$ and $\pi(\sigma)$ given by (1.3.22) and (1.4.1), and the conditional posteriors $\pi(\mu|y,\sigma)$ and $\pi(\sigma|\mu,y)$ given by (2.6.4) and (2.6.7). It can then be shown (book 2, p. 19) that the marginal likelihood is approximated by:

$$f(y) \approx \pi^{-n/2} \, \frac{\delta^{\alpha/2}}{\bar{\delta}^{\bar{\alpha}/2}} \, \frac{\Gamma(\bar{\alpha}/2)}{\Gamma(\alpha/2)} \, \frac{\exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v}\right)}{\frac{1}{J}\sum_{j=1}^{J}(1+vn/\sigma)^{1/2}\exp\left(-\frac{1}{2}\frac{(\mu-\bar{m})^2}{\bar{v}}\right)} \tag{2.6.17}$$

The expression is evaluated at the high density points μ^* and σ^*, taken to be the median of the Gibbs sampler draws for the posterior. It ressembles much (1.4.17), except for the final term which represents the Gibbs sampler approximation. Using (2.6.17), we find $m(y) = -29.12$. The value is consistent with the value of -26.07 obtained in section 4.6. Jeffrey's guidelines (Table 4.1) suggest decisive support in favor of the hierarchical prior model: the independent prior model is not the one most supported by the data.

CHAPTER 7

The Metropolis-Hastings algorithm

7.1 Metropolis-Hastings: motivation

Consider again the stock return example introduced in chapter 3, but assume we adopt a slightly different formulation. The return is still modelled as a normal distribution with mean μ, but the variance is now expressed as $\exp(\lambda)$, with λ a real-valued parameter. The exponential guarantees that whatever the value of λ, the variance will always be positive. In this model, the parameters of interest are thus $\theta = \{\mu, \lambda\}$.

Denoting by y_i the stock return on year i, we have $f(y_i) \sim N(\mu, \exp(\lambda))$ and the probability density function for each return is given by:

$$f(y_i|\mu,\lambda) = (2\pi \ \exp(\lambda))^{-1/2} \ \exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{\exp(\lambda)}\right) \tag{2.7.1}$$

Using definition 3.4, the likelihood function then obtains as:

$$f(y|\mu,\lambda) = (2\pi \ \exp(\lambda))^{-n/2} \ \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\exp(\lambda)}\right) \tag{2.7.2}$$

For the prior, we follow as usual definition 4.1 and assume independence between μ and λ so that $\pi(\mu,\lambda) = \pi(\mu)\pi(\lambda)$. The prior distribution for μ is unchanged: $\pi(\mu) \sim N(m,v)$. it is thus given by (1.3.22):

$$\pi(\mu) = (2\pi v)^{-1/2} \ \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v}\right) \tag{2.7.3}$$

We then need a prior for λ. Because λ can take any real value, we choose again a normal distribution so that $\pi(\lambda) \sim N(g,z)$ with g the prior mean and z the prior variance. Following:

$$\pi(\lambda) = (2\pi z)^{-1/2} \ \exp\left(-\frac{1}{2}\frac{(\lambda-g)^2}{z}\right) \tag{2.7.4}$$

Applying then Bayes rule 3.3, we obtain:

$$\pi(\mu,\lambda|y) \propto \exp(\lambda)^{-n/2} \ \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\exp(\lambda)}\right) \times \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v}\right) \times \exp\left(-\frac{1}{2}\frac{(\lambda-g)^2}{z}\right) \tag{2.7.5}$$

As usual, any multiplicative term not involving μ or λ has been relegated to the normalization constant. This is a joint posterior distribution that cannot be marginalized analitycally. We first try to calculate the conditional posterior distributions in order to use the Gibbs sampling algorithm. Consider the conditional posterior $\pi(\mu|y,\lambda)$. Using definition 6.1, we start from (2.7.5) and relegate to the normalization constant any term not involving μ. This yields:

$$\pi(\mu|y,\lambda) \propto \ \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\exp(\lambda)}\right) \times \exp\left(-\frac{1}{2}\frac{(\mu-m)^2}{v}\right) \tag{2.7.6}$$

Rearranging and completing the squares eventually yields (book 2, p. 19):

$$\pi(\mu|y,\lambda) \propto \exp\left(-\frac{1}{2}\frac{(\mu-\bar{m})^2}{\bar{v}}\right) \tag{2.7.7}$$

with:

$$\bar{v}=\left(\frac{n}{\exp(\lambda)}+\frac{1}{v}\right)^{-1} \qquad \bar{m}=\bar{v}\left(\frac{1}{\exp(\lambda)}\sum_{i=1}^{n}y_i+\frac{m}{v}\right) \tag{2.7.8}$$

This is the kernel of a normal distribution with mean $\bar{m}$ and variance $\bar{v}$: $\pi(\mu|y,\lambda) \sim N(\bar{m},\bar{v})$.

Consider now the conditional posterior $\pi(\lambda|y,\mu)$. Using definition 6.1, we start from (2.7.5) and relegate to the normalization constant any term not involving λ. This yields:

$$\pi(\lambda|y,\mu) \propto \exp(\lambda)^{-n/2}\ \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\exp(\lambda)}\right)\times\exp\left(-\frac{1}{2}\frac{(\lambda-g)^2}{z}\right) \tag{2.7.9}$$

This is a complex expression in λ that cannot be rearranged into a known distribution and is thus intractable. Even though we managed to calculate the conditional posterior, it is of unknown form and thus cannot be used for the Gibbs sampling algorithm. In this case we need a more general approach, which is given by the Metropolis-Hastings algorithm.

7.2 Metropolis-Hastings: the algorithm

Consider a model with n parameters so that $\theta = \{\theta_1,\cdots,\theta_n\}$. Assume that the conditional posteriors $\pi(\theta_1|y,\theta_2,\cdots,\theta_n)$, $\cdots$, $\pi(\theta_n|y,\theta_1\cdots,\theta_{n-1})$ can be calculated, but that for at least one parameter (say θ_i) this posterior is non-standard so that it is not possible to sample values directly from $\pi(\theta_i|y,\theta_1,\cdots,\theta_n)$. In this case, we can use the Metropolis-Hastings algortihm. Unlike the Gibbs sampling algorithm where new values are obtained at each iteration, The Metropolis-Hastings algorithm will only generate candidate values, and accept them with a certain probability. If the draw is rejected, the value inherited from the previous iteration is retained.

Concretely, the Metropolis-Hastings first requires a function that produces a candidate value for the current iteration, given the previous iteration value.

definition 7.1: let $\theta_i^{(j)}$ denote the value of θ_i at iteration j; a **transition kernel** is a probability density function $q(\theta_i^{(j-1)},\theta_i^{(j)})$ for $\theta_i^{(j)}$ with respect to the value $\theta_i^{(j-1)}$.

Some common choices of transition kernels are the random walk kernel and the independence kernel. The **random walk kernel** is of the form:

$$\theta_i^{(j)}=\theta_i^{(j-1)}+x \tag{2.7.10}$$

where x is a random variable with known distribution, for instance $\pi(x) \sim N(0,\tau)$, with τ a user-specified variance term defining the amplitude of the move. The **independence kernel** is defined as:

$$\theta_i^{(j)}=x \tag{2.7.11}$$

where x is a random variable with known distribution. In this case, at every iteration j a value $\theta_i^{(j)}$ is sampled directly from the candidate distribution independently of the previous value $\theta_i^{(j-1)}$, hence the name independence kernel.

Once a suitable transition kernel is chosen, it remains to determine the probability of acceptance of the candidate.

definition 7.2: let $\pi(\theta_i|y,\theta_1,\cdots,\theta_n)$ denote the conditional posterior for θ_i, and $q(\theta_i^{(j-1)},\theta_i^{(j)})$ denote the transition kernel; the **probability of acceptance** is the function $\alpha(\theta_i^{(j-1)},\theta_i^{(j)})$ given by:

$$\alpha(\theta_i^{(j-1)},\theta_i^{(j)}) = \min\left\{1, \frac{\pi(\theta_i^{(j)}|y,\theta_1,\cdots,\theta_n)\ \ q(\theta_i^{(j)},\theta_i^{(j-1)})}{\pi(\theta_i^{(j-1)}|y,\theta_1,\cdots,\theta_n)\ \ q(\theta_i^{(j-1)},\theta_i^{(j)})}\right\}$$

Roughly speaking, the move is accepted with probability 1 if the density of the candidate value $\pi(\theta_i^{(j)}|y,\theta_1,\cdots,\theta_n)\ \ q(\theta_i^{(j)},\theta_i^{(j-1)})$ is higher than that inherited from the previous iteration $\pi(\theta_i^{(j-1)}|y,\theta_1,\cdots,\theta_n)\ q(\theta_i^{(j-1)},\theta_i^{(j)})$. Conversely, if the density of the candidate value is lower, the candidate will only be accepted with a probability smaller than 1.

The Metropolis-Hastings algorithm can then be summarized as follows:

algorithm 7.1: Metropolis-Hastings algorithm

1. set any initial values $\theta_i^{(0)}$ for θ_i.
2. at iteration j, obtain a candidate value $\tilde{\theta}_i$ from $q(\theta_i^{(j-1)},\theta_i^{(j)})$.
3. determine the probability of acceptance from $\alpha(\theta_i^{(j-1)},\theta_i^{(j)})$.
4. draw a uniform random number u from $u \sim U(0,1)$.
5. if $u \leq \alpha(\theta_i^{(j-1)},\theta_i^{(j)})$, accept the candidate and set $\theta_i^{(j)} = \tilde{\theta}_i$; else, reject the candidate and set $\theta_i^{(j)} = \theta_i^{(j-1)}$.
6. repeat until the desired number of iterations is realised.

The choice of a transition kernel represents a key feature of the algorithm. The random walk and independence kernels are simple, but not necessarily optimal choices. Ideally, a good kernel should allow for sufficient variability in the value of θ_i between two iterations. This ensures that a large part of the support of $\pi(\theta_i|y)$ will be covered by the iterations of the algorithm, which improves the mixing between iterations and the quality of the empirical posterior. However, larger differences between $\theta_i^{(j)}$ and $\theta_i^{(j-1)}$ typically imply larger differences between $\pi(\theta_i^{(j)}|y,\theta_1,\cdots,\theta_n)$ and $\pi(\theta_i^{(j-1)}|y,\theta_1,\cdots,\theta_n)$, which increases the probability of rejection. Then some values may be repeated often, resulting in a poor empirical distribution. The kernel must thus be chosen to generate the most efficient compromise between these two aspects, and this is usually achieved by calibrating it to produce an acceptance rate somewhere around 20-30%.

Whatever the acceptance rate, the Metropolis-Hastings algorithm is constructed to produce repeated values. To avoid an empirical distribution that is too coarse it is customary to discard a fraction of the draws, retaining only every n draws, where n is for instance 10 or 20. This technique is known as **thinning**. It effectively solves the issue of repeated values but multiplies by n the total number of draws to compute. Following, the computational cost of the Metropolis-Hastings algorithm increases dramatically.

Finally, it is worth noting that the Metropolis-Hastings algorithm can be integrated to a standard Gibbs sampling framework. If $\theta_1,\cdots,\theta_n$ are the parameters of interest and only θ_i has a non-standard distribution, then θ_i can be simulated from Metropolis-Hastings while the other parameters are obtained from the Gibbs sampling methodology.

7.3 Metropolis-Hastings: an example

We now return to our stock return example. As shown by equation (2.7.7), the conditional posterior $\pi(\mu|y,\lambda)$ is standard: $\pi(\mu|y,\lambda) \sim N(\bar{m},\bar{v})$. It can thus be sampled directly from the Gibbs sampling algorithm. On the other hand, the conditional posterior distribution $\pi(\mu,\lambda|y)$ given by (2.7.9) is non-standard and requires the Metropolis-Hastings algorithm. First, define a transition kernel for λ. Here the simple random walk kernel is chosen:

$$\lambda^{(j)} = \lambda^{(j-1)} + x \qquad \pi(x) \sim N(0,\tau) \tag{2.7.12}$$

It follows that $q(\lambda^{(j-1)},\lambda^{(j)}) \sim N(\lambda^{(j-1)},\tau)$. Also, (2.7.12) and the symmetry of $\pi(x)$ around 0 implies that $q(\lambda^{(j)},\lambda^{(j-1)}) \sim N(\lambda^{(j)},\tau)$. Following, we conclude that $q(\lambda^{(j-1)},\lambda^{(j)}) = q(\lambda^{(j)},\lambda^{(j-1)})$, which conveniently simplifies the probability of acceptance in defintion 7.2 to:

$$\alpha(\lambda^{(j-1)},\lambda^{(j)}) = \min\left\{1, \frac{\pi(\lambda^{(j)}|y,\mu)}{\pi(\lambda^{(j-1)}|y,\mu)}\right\} \tag{2.7.13}$$

Given (2.7.9), this directly yields (book 2, p. 22):

$$\begin{aligned} &\alpha(\lambda^{(j-1)},\lambda^{(j)}) \\ = \;& \min\left\{1, \exp\left(\frac{1}{2}\left[\begin{array}{l} n(\lambda^{(j-1)}-\lambda^{(j)}) + \left[\exp(-\lambda^{(j-1)}) - \exp(-\lambda^{(j)})\right]\sum_{i=1}^{n}(y_i-\mu)^2 \\ +\dfrac{(\lambda^{(j-1)}-g)^2-(\lambda^{(j)}-g)^2}{z} \end{array}\right]\right)\right\} \end{aligned} \tag{2.7.14}$$

Following, the algorithm for the model obtains as:

algorithm 7.2: Gibbs sampling/Metropolis-Hastings algorithm for the stock return model

1. set initial values $\mu^{(0)}$ and $\lambda^{(0)}$; use the prior means $\mu^{(0)} = m$ and $\lambda^{(0)} = g$.

2. at iteration j:

   draw $\mu^{(j)}$ from $\pi(\mu|y,\lambda^{(j-1)}) \sim N(\bar{m},\bar{v})$ with:

   $$\bar{v} = \left(\frac{n}{\exp(\lambda)^{(j-1)}} + \frac{1}{v}\right)^{-1} \qquad \bar{m} = \bar{v}\left(\frac{1}{\exp(\lambda^{(j-1)})}\sum_{i=1}^{n} y_i + \frac{m}{v}\right)$$

3. at iteration j:

   draw a candidate $\tilde{\lambda}$ from $\tilde{\lambda} = \lambda^{(j-1)} + x \quad , \quad \pi(x) \sim N(0,\tau)$

4. at iteration j: obtain the acceptance probability $\alpha(\lambda^{(j-1)},\lambda^{(j)})$ given by:

   $$\min\left\{1, \exp\left(\tfrac{1}{2}\left[\begin{array}{l} n(\lambda^{(j-1)}-\lambda^{(j)}) + \left[\exp(-\lambda^{(j-1)}) - \exp(-\lambda^{(j)})\right]\sum_{i=1}^{n}(y_i-\mu)^2 \\ +\dfrac{(\lambda^{(j-1)}-g)^2-(\lambda^{(j)}-g)^2}{z} \end{array}\right]\right)\right\}$$

5. at iteration j: draw a uniform random number u from $u \sim U(0,1)$.

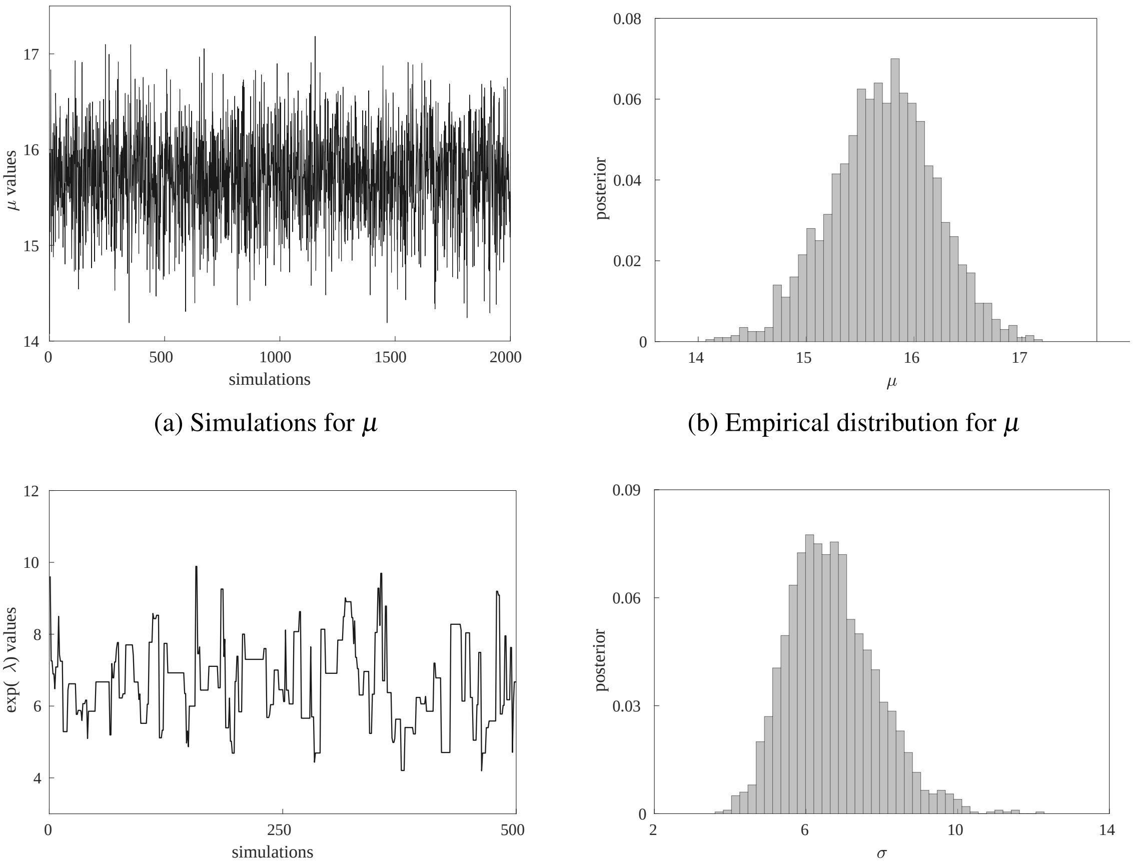
   if $u \leq \alpha(\theta_i^{(j-1)},\theta_i^{(j)})$, set $\theta_i^{(j)} = \tilde{\theta}_i$; else, set $\theta_i^{(j)} = \theta_i^{(j-1)}$

6. repeat to obtain 1000 iterations as burn-in sample and 2000 additional iterations for simulated values.

It remains to calibrate the prior $\pi(\lambda)$ and the transition kernel $q(\lambda^{(j-1)}, \lambda^{(j)})$. For $\pi(\lambda)$, we set $g = 1.6$ and $z = 0.04$. This way, the mean and variance of $\exp(\lambda)$ match the prior mean of 5 and the prior variance of 1 proposed for σ in section 4.3. For the random walk kernel $q(\lambda^{(j-1)}, \lambda^{(j)})$ we set $\tau = 0.5$, which results in an acceptance rate of roughly 25%.

The algorithm is then run for 1000 burn-in iterations and 2000 samples, multiplied by 20 to retain only every 20 simulated value. The resulting simulated values along with the associated empirical distributions are displayed in Figure 7.1. The top panels show the simulations obtained for the Gibbs sampling step for μ along with the resulting empirical distribution. These plots are quite consistent with the top plots in Figure 6.1.

The bottom plots display the simulations and empirical distribution from the Metropolis-Hastings algorithm for λ. The left panel shows the first 500 iterations of the algorithm, before trimming is operated. The repeated values typical of the Metropolis-Hastings algorithm are quite apparent. The right panel displays the empirical distribution obtained after posterior trimming. The distribution looks quite smooth, demonstrating the gain in accuracy from trimming. It is consistent with the distribution in Figure 6.1 though slightly tighter, a feature resulting from the alternative formulation of the model.



(a) Simulations for μ

(b) Empirical distribution for μ

(c) Simulations for $\exp(\lambda)$ (500 iterations)

(d) Empirical distribution for $\exp(\lambda)$

Figure 7.1: Gibbs sampling simulations and empirical distributions for μ and $\exp(\lambda)$

7.4 Marginal likelihood with Metropolis-Hastings

Section 6.5 introduced the Chib (1995) method to calculate the approximate marginal likelihood whenever sampling from the Gibbs algorithm is available. Chib and Jeliazkov (2001) propose an adaptation of the methodology to the Metropolis-Hastings algorithm. However, their approach is more complicated. Also, for both approaches the computational cost may become prohibitive when the model involves more than two parameters. For this reason, we introduce here the simpler and more general methodology of Gelfand and Dey (1994). The approach is conceptually simple and relies on an harmonic mean approximation. It only requires simulated draws from the marginal posteriors, regardless of the method used to produce them.

Consider any probability density function $g(\theta)$. Then we have the following identity:

$$\mathbb{E}\left(\frac{g(\theta)}{\pi(\theta)\,f(y|\theta)}\middle|y\right)=\frac{1}{f(y)} \tag{2.7.15}$$

Indeed, it is immediate that:

$$\mathbb{E}\left(\frac{g(\theta)}{f(y|\theta)\,\pi(\theta)}\middle|y\right)=\int\frac{g(\theta)}{f(y|\theta)\,\pi(\theta)}\pi(\theta|y)d\theta=\int\frac{g(\theta)}{f(y|\theta)\,\pi(\theta)}\frac{f(y|\theta)\pi(\theta)}{f(y)}d\theta=\frac{1}{f(y)}\int g(\theta)d\theta=\frac{1}{f(y)} \tag{2.7.16}$$

In practice, the expectation is unknown. However, a consistent estimate can be obtained from the Gibbs sampler values, yielding the following approximation:

$$\frac{1}{f(y)}\approx\frac{1}{J}\sum_{j=1}^{J}\frac{g(\theta^{(j)})}{f(y|\theta^{(j)})\,\pi(\theta^{(j)})} \tag{2.7.17}$$

In theory, any probability density function $g(\theta)$ can be used to compute the approximation. In practice, the choice of $g(\theta)$ is very important for the accuracy of the approximation. Geweke (1999) propose to use a truncated multivariate normal distribution: $g(\theta)\sim\bar{N}(\hat{\theta},\hat{\Sigma})$, where $\hat{\theta}$ and $\hat{\Sigma}$ denote the empirical posterior moments of the model parameters, calculated as:

$$\hat{\theta}=\frac{1}{J}\sum_{j=1}^{J}\theta^{(j)}\qquad\qquad\hat{\Sigma}=\frac{1}{J}\sum_{j=1}^{J}(\theta^{(j)}-\hat{\theta})(\theta^{(j)}-\hat{\theta})' \tag{2.7.18}$$

The truncation is realised through the region $\hat{\Theta}=\{\theta:(\theta-\hat{\theta})'\hat{\Sigma}^{-1}(\theta-\hat{\theta})\leq\chi^2_{1-\omega}(k)\}$, where $\chi^2_{1-\omega}(k)$ is the $1-\omega$ quantile of the Chi-squared distribution with k degrees of freedom, for k the dimension of θ and $\omega\in[0,1]$ some probability set by the statistician. We then obtain:

$$g(\theta)=\omega^{-1}(2\pi)^{-k/2}|\hat{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\theta-\hat{\theta})'\hat{\Sigma}^{-1}(\theta-\hat{\theta})\right)\mathbb{1}(\theta\in\hat{\Theta}) \tag{2.7.19}$$

where $\mathbb{1}(\theta\in\hat{\Theta})$ is the indicator function equal to 1 if θ is in $\hat{\Theta}$, and 0 otherwise. The function thus truncates the extreme values of θ that may result in imprecise estimates of (2.7.17). Common choices for ω are $\omega=0.5$, $\omega=0.25$ and $\omega=0.1$.

We now apply this method to the stock return example. Given $\theta=\{\mu,\lambda\}$, (2.7.17) becomes:

$$\frac{1}{f(y)}\approx\frac{1}{J}\sum_{j=1}^{J}\frac{g(\theta^{(j)})}{f(y|\mu^{(j)},\lambda^{(j)})\,\pi(\mu^{(j)})\,\pi(\lambda^{(j)})} \tag{2.7.20}$$

Using the density (2.7.19) along with the likelihood function (2.7.2) and the priors (2.7.3) and (2.7.4), we obtain (book 2, p. 23):

$$\begin{aligned}\frac{1}{f(y)} &\approx (\omega J)^{-1}(2\pi)^{n/2}|\hat{\Sigma}|^{-1/2}(vz)^{1/2} \\ &\quad \times \sum_{j=1}^{J} \mathbb{1}(\theta \in \hat{\Theta}) \times \exp\left(\frac{1}{2}\left[n\lambda + \sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\exp(\lambda)} + \frac{(\mu-m)^2}{v} + \frac{(\lambda-g)^2}{z} - (\theta-\hat{\theta})'\hat{\Sigma}^{-1}(\theta-\hat{\theta})\right]\right)\end{aligned} \tag{2.7.21}$$

Applying (2.7.21) with $\omega = 0.5$, we find $m(y) = -29.10$. This value is virtually equal to the marginal likelihood found in section 6.5 for the independent prior model. This indicates that the two models are equally supported by the data and are, in fact, extremely similar.

CHAPTER 8

Mathematical theory

This chapter introduces the mathematical foundations behind the Gibbs sampling and Metropolis-Hastings methodologies. The chapter is technical and may safely be skipped if one is interested in methods only. A good treatment of the subject can be found in Chib and Greenberg (1995) and Greenberg (2012) , chapters 6-7. The presentation in this part follows more or less the same guidelines.

8.1 Markov Chains with finite state space

Assume our objective is to sample values from some target statistical distribution. For the time being we keep things simple and assume that the distribution takes values in the finite set $S = \{s_1, \cdots, s_n\}$. Consider for instance a random variable taking values in $S = \{1,2\}$ with $f(1) = 0.4$ and $f(2) = 0.6$, as shown by Figure 8.1.

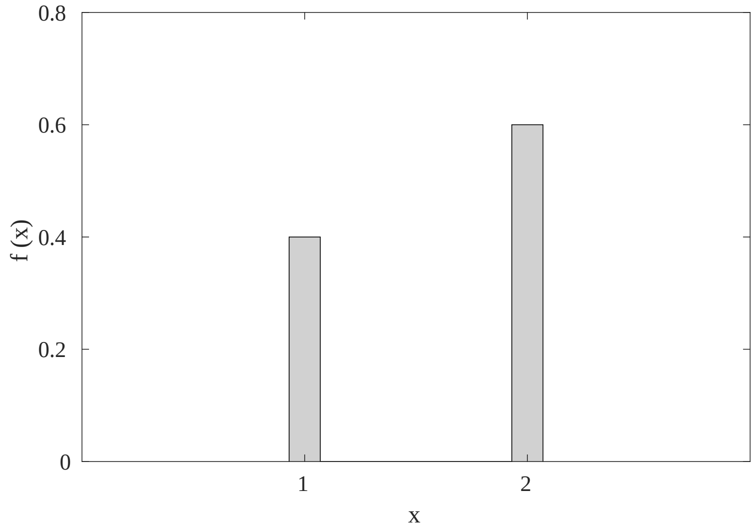


Figure 8.1: Probability mass function of the target distribution

To generate draws from this distribution, we will use **Markov chains**, a type of stochastic processes. Consider for instance a discrete time stochastic process X_t which takes values in $S = \{s_1, \cdots, s_n\}$ (similarly to the target distribution), with $t = 1, 2, \cdots$. The stochastic process is then just a collection of random variables $X_1, X_2, \cdots$. The n possible values of X_t are called the **states** of the system, and we are interested in describing the probabilities that the process moves from one state to another over a period of time. Concretely, for $s_i, s_j \in S$, we call the **transition probabilities** the set of values p_{ij} such that $p_{ij} = \mathbb{P}(X_{t+1} = s_j | X_t = s_i)$.

definition 8.1: a **finite Markov chain** is a stochastic process X_t with states $S = \{s_1, \cdots, s_n\}$ and transition probabilities $p_{ij} = \mathbb{P}(X_{t+1} = s_j | X_t = s_i), \quad \text{for } s_i, s_j \in S.$

Whenever p_{ij} does not depend on t, we say that the Markov chain is **homogenous**. In this case, the dynamics of the process can be conveniently summarized in a single matrix known as the transition matrix.

definition 8.2: the **transition matrix** is the $n \times n$ matrix $P = \{p_{ij}\}$ such that $p_i = (p_{i1}, \cdots, p_{in})$ is row i of P, and $\sum_{j=1}^{n} p_{ij} = 1$.

For instance, consider the simple homogenous Markov chain X_t with states $S = \{1, 2\}$ and transition matrix:

$$P = \begin{pmatrix} 2/3 & 1/3 \\ 2/9 & 7/9 \end{pmatrix} \tag{2.8.1}$$

P says that while in state 1 at period t, the probability to remain in state 1 at period $t+1$ is $2/3$ while the probability to move to state 2 is $1/3$. Starting from state 2 at period t, the probability to move to state 1 at period $t+1$ is $2/9$ while the probability to stay in state 2 is $7/9$.

We now want to determine the probability $p_{ij}^{(2)}$ to move from state s_i at period t to state s_j at period $t+2$. To do so, we first need to move from state s_i to some state s_k during the first period, then move from state s_k to state s_j during the second period, for any $s_k \in S$. In other words, $p_{ij}^{(2)} = \sum_{k=1}^{n} p_{ik} p_{kj}$. It can be verified that this implies $P^{(2)} = PP = P^2$. Working by induction, we then obtain that $P^{(h)} = P^h$. For example:

$$P^{(3)} = P^3 = \begin{pmatrix} 0.452 & 0.548 \\ 0.364 & 0.636 \end{pmatrix} \tag{2.8.2}$$

P^3 says that the probability to move from state 1 at period t to state 2 at period $t+3$ is 0.548, while the probability to return to state 1 is 0.452.

Typically, we are interested in $P^{(h)}$ when h gets large. Table 8.1 reports the transition probabilities for different horizons h.

period (h)	$p_{11}^{(h)}$	$p_{12}^{(h)}$	$p_{21}^{(h)}$	$p_{22}^{(h)}$
1	0.667	0.333	0.222	0.778
2	0.518	0.482	0.321	0.679
3	0.453	0.547	0.365	0.635
4	0.423	0.577	0.384	0.616
5	0.410	0.590	0.393	0.607
10	0.401	0.599	0.399	0.601
20	0.400	0.600	0.400	0.600

Table 8.1: Transition probabilities for P at different horizons

The matrix entries converge to some equilibrium values. In matrix form, we find that:

$$\lim_{h \to +\infty} P^{(h)} = \begin{pmatrix} 0.400 & 0.600 \\ 0.400 & 0.600 \end{pmatrix} \tag{2.8.3}$$

We observe that the rows of the long-term matrix are similar: for h large enough, the probability of being in state s_j at period $t+h$ is the same, whatever the state s_i we start at period t. This remarkable property constitutes the foundation of modern simulation methods, and is related to the notion of invariant distribution.

definition 8.3: let P be a transition matrix; the probability vector $\pi = (\pi_1, \cdots, \pi_n)$ is an **invariant distribution** if:

$\pi' P = \pi$

This definition says that if we select states at period t with probabilities π then move to period $t+1$ according to P (left-hand side), the states at $t+1$ will still be drawn according to π (right-hand side). Following, π represents the state probabilities of the Markov chain at any time, hence the name invariant distribution.

Let us compute the invariant distribution for the transition matrix P in (2.8.2). From definition 8.3, we obtain:

$$\begin{pmatrix}\pi_1 & \pi_2\end{pmatrix}\begin{pmatrix}2/3 & 1/3 \\ 2/9 & 7/9\end{pmatrix} = \begin{pmatrix}\pi_1 & \pi_2\end{pmatrix} \tag{2.8.4}$$

The first row yields $2/3\pi_1 + 2/9\pi_2 = \pi_1$. Using then $\pi_2 = 1 - \pi_1$ and solving for π_1 yields $\pi_1 = 2/5$, and $\pi_2 = 3/5$. Thus $\pi = (0.4,\ 0.6)$ which corresponds to the rows of the long-term matrix in equation (2.8.3). In other words, after a sufficient number of periods h, the Markov chain converges to the invariant distribution $\pi = (0.4,\ 0.6)$. Conveniently, this invariant distribution corresponds to the target distribution depicted in Figure 8.1.

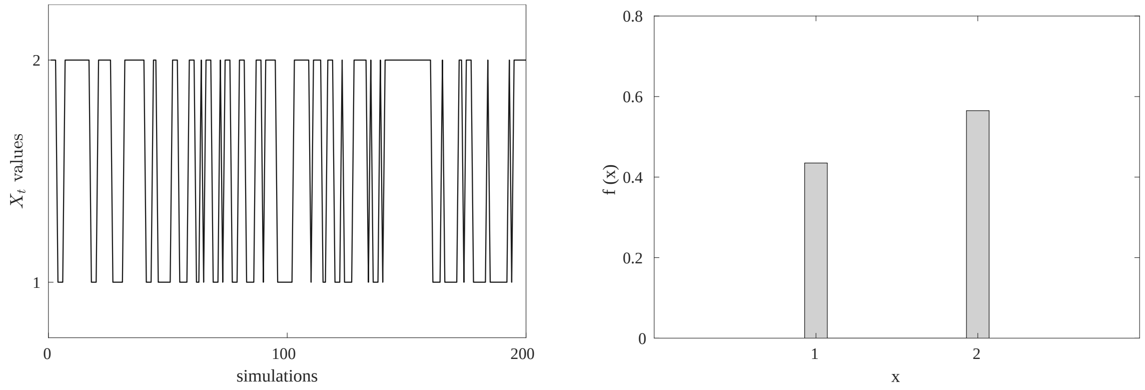
This suggests a natural procedure to generate values from a target finite distribution:

algorithm 8.1: distribution sampling with finite Markov chain

1. create a finite Markov chain with transition matrix P such that the invariant distribution corresponds to the target distribution.
2. set any state as the initial state X_0 of the Markov chain.
3. run the Markov chain for h periods; that is, determine $X_1, \cdots, X_h$, randomly moving from period t to period $t+1$ according to the transition matrix P.
4. for h large enough, the Markov chain has reached the invariant distribution; run the Markov chain for an additional k periods, that is, determine $X_{h+1}, \cdots, X_{h+k}$ according to P.
5. discard $X_1, \cdots, X_h$; then $X_{h+1}, \cdots, X_{h+k}$ are drawn from the invariant distribution, which corresponds by construction to the target distribution.

Table 8.1 makes it clear why the initial values $X_1, \cdots, X_h$ must be discarded. For early periods the invariant distribution is not yet reached, and the state of the Markov chain still depends significantly on the initial state. It is thus important to run the chain for sufficiently long and to clear the influence of the initial state.

The use of algorithm 8.1 is illustrated in Figure 8.2. We use the transition matrix P defined in equation (2.8.1) and run the Markov chain for 250 periods, setting the initial state as 1. The first 50 periods are discarded as burn-in sample, which is sufficient to reach the invariant distribution of the chain, as shown in Table 8.1. The empirical distribution resulting from the chain is quite close to the target distribution shown in Figure 8.1. Because only 200 values are sampled, the empirical distribution does not replicate exactly the target distribution, but the approximation could be made arbitrarily accurate by increasing the number of observations generated.



(a) Markov chain values X_t (b) Empirical distribution of X

Figure 8.2: Distribution sampling with finite state Markov chain

Given a finite Markov chain and the associated transition matrix P, is it always possible to converge to an invariant distribution? And if yes, is this invariant distribution unique? To answer these questions we first need a few definitions, starting with the notion of communicating states.

definition 8.4: let X_t be a finite Markov chain with states $S = \{s_1, \cdots, s_n\}$; we say that state s_j is **reachable** from s_i, denoted by $s_i \to s_j$, if there is some $h \geq 1$ with $p_{ij}^{(h)} > 0$.

If state s_i is reachable from s_j and state s_j is reachable from s_i, we say that states s_i and s_j **communicate**, denoted by $s_i \leftrightarrow s_j$.

Basically, two states communicate if from one, it is possible to reach the other at some point. For instance, consider the Markov chain with transition matrix:

$$Q = \begin{pmatrix} 2/3 & 1/3 \\ 0 & 1 \end{pmatrix} \tag{2.8.5}$$

We can see that states 1 and 2 don't communicate: if we ever reach state 2, we will remain in it forever and so state 1 is not reachable from state 2. An important class of Markov chains is that where all the states communicate.

definition 8.5: a Markov chain is **irreducible** if all states communicate.

Another important property of Markov chains is periodicity.

definition 8.6: let X_t be a finite Markov chain with states $S = \{s_1, \cdots, s_n\}$; state $s_j \in S$ is **periodic** of period d if there exists some integer $d \geq 1$ such that $p_{jj}^{(h)} > 0$ whenever h is a multiple of d, and $p_{jj}^{(h)} = 0$ otherwise. The chain is **aperiodic** if the period is 1 for all the states.

Simply speaking, a state has period d if it takes a multiple of d periods to return to it. Consider for instance the Markov chain with transition matrix:

$$R = \begin{pmatrix} 2/3 & 1/3 \\ 1 & 0 \end{pmatrix} \tag{2.8.6}$$

Whenever the chain is in state 2, it can only move to state 1. Returning to state 2 thus takes at least two periods: one to move to state 1, and one to reach state 2 again. State 2 has thus a period of 2, and the chain is not aperiodic.

It is now possible to state the main result of this section:

theorem 8.1: let X_t be an irreducible and aperiodic Markov chain over the finite states $S = \{s_1, \cdots, s_n\}$; then there exists a unique probability distribution π such that $\pi' P = \pi'$; also:

$|p_{ij}^{(h)} - \pi_j| \leq \delta^{\,h/v} \qquad$ for all $i, j = 1, \cdots, n$

with $0 < \delta < 1$ and v some positive integer.

This theorem lies at the basis of Monte Carlo Markov Chain (MCMC) methods. In a finite state space, it says that as long as we can define a Markov chain that is irreducible and aperiodic, there exists for sure a unique invariant distribution for the chain. Also, for sufficiently large h, the chain converges to the invariant distribution π at some geometric rate h/v.

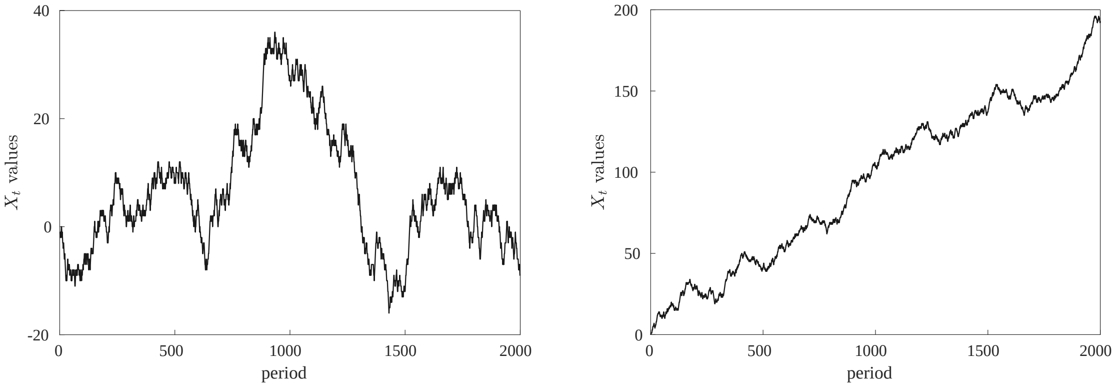
Understanding why the Markov chain has to be irreducible and aperiodic is straightforward. If the chain is not irreducible, then the exist at least two states s_i and s_j that don't communicate. In this case it is not possible to reach an invariant distribution since reaching state s_i precludes state s_j to be ever joined later on. If the chain is not aperiodic, there exists at least one state s_j such that $p_{jj}^{(h)} > 0$ whenever h is a multiple of d, and $p_{jj}^{(h)} = 0$ otherwise. Thus by definition we cannot have $p_{jj}^{(h)} = \pi_j$ for all periods.

8.2 Markov Chains with countable state space

Markov chains with finite states prove often too restrictive for empirical applications. As a first generalization we consider Markov chains with countable state spaces. Such Markov chain take an infinite, but still countable number of values. A classical example is the random walk process with states $S = \mathbb{Z}$, the set of integers, and transition probabilities given by:

$$p_{ij} = \begin{cases} p, & \text{if } j = i+1 \\ q, & \text{if } j = i \\ r, & \text{if } j = i-1 \end{cases} \qquad p + q + r = 1 \tag{2.8.7}$$

Whenever the state space S is countable, irreducibility and aperiodicity are not sufficient anymore to guarantee the existence of a unique invariant distribution. To see this, consider Figure 8.3.



(a) $p = q = r = 1/3$

(b) $p = 0.4$, $q = 0.25$, $r = 0.35$

Figure 8.3: Examples of random walk processes

The two processes are obtained from the Markov chain (2.8.7). The process on the left obtains from $p = q = r = 1/3$ while that on the right is generated with $p = 0.4$, $q = 0.25$, $r = 0.35$. The right process is an example of a biased Markov chain where the probability to move up larger than the probability to move down.

Clearly, both processes are irreducible and aperiodic. The process on the left looks stationary. However, the right process is drifting off to infinity due to its bias. Therefore in the long run the probability to reach any finite value s_i tends to 0: $p_{ij}^{(h)} \to 0$ for all i, j. Because the probabilities of reaching finite states decline over time, the process cannot converge to an invariant distribution where the probability to obtain any state s_i remains constant over periods.

We thus need a stronger concept, which is the notion of recurrence. This first requires a few definitions.

definition 8.7: let X_t be a Markov chain with countable states $S = \{s_1, s_2, \cdots\}$, and let $X_0 = s_i$; the **return time** T_i is the number of periods for the chain to first return to s_i:

$T_i = min\{t \geq 1 : X_h = s_i\}$

The return time T_i is a random variable. For instance, for the Markov chain defined in equation (2.8.7), we have $T_i = 1$ with probability q, $T_i = 2$ with probability $2pr$ (the chain moves up then down, or the converse), and so on. Formally, we denote the probability of return time at period h by $f_i^{(h)} = \mathbb{P}(T_i = h | X_0 = s_i)$. From this, the probability of ever returning to s_i is given by:

$$f_i = \sum_{h=1}^{\infty} f_i^{(h)} \tag{2.8.8}$$

We can then define the concept of recurrence.

definition 8.8: let f_i denote the probability of returning to s_i; the state s_i is **recurrent** if $f_i = 1$; otherwise, s_i is **transient** if $f_i < 1$.

Basically, a state is recurrent if the chain returns to it at some point with probability 1. Certainly, a chain cannot reach an invariant distribution if some of its states are transient. Recurrence, however, is not sufficient to guarantee a unique invariant distribution. For a state s_i, define the mean return time m_i as:

$$m_i = \mathbb{E}(T_i | X_0 = s_i) = \sum_{h=1}^{\infty} h \, f_i^{(h)} \tag{2.8.9}$$

We then define positive recurrence as:

definition 8.9: let m_i denote the mean return time to s_i; the state s_i is **positive recurrent** if $m_i < \infty$; otherwise, s_i is **null recurrent** if $m_i = \infty$.

A state is positive recurrent if returning to it takes on average a finite number of periods only. It is null recurrent if returning to it happens with probability 1, but takes on average an infinite number of periods. With these elements, it is possible to define the conditions under which a unique invariant distribution is guaranteed.

theorem 8.2: let X_t be an irreducible Markov chain with countable states $S = \{s_1, s_2, \cdots\}$; then:

1. if all states are recurrent, they are either all positive recurrent or all null recurrent.

2. there exists an invariant distribution if and only if all states are positive recurrent; in this case, the invariant distribution $\pi = (\pi_1, \pi_2, \cdots)$ is unique and given by:

$\pi_i = 1/m_i \qquad$ for all $s_i \in S$

3. If the states are positive recurrent, then $\pi_i = \lim\limits_{h \to +\infty} 1/h \sum\limits_{t=1}^{h} \mathbb{1}(X_t = s_i)$

The first part of the theorem states that an irreducible Markov chain will either return to all states within finite mean times, or to none of them. The second part says that the existence of an invariant distribution is equivalent to all the states being positive recurrent, in which case the invariant distribution is the inverse of the mean return time for each state. The final part provides a way to recover the distribution from the empirical frequency of each state s_i, provided the number of observations h is sufficiently large.

Theorem 8.2 provides a measure of convergence for irreducible Markov chains in time average. To obtain instead convergence from transition probabilities, in the sense that $\pi_j = \lim\limits_{h \to +\infty} p_{ij}^{(h)}$ and regardless of the initial state s_i, then the further condition of aperiodicity is needed on the chain. We have the following theorem:

theorem 8.3: let X_t be an irreducible Markov chain with invariant distribution $\pi = (\pi_1, \pi_2, \cdots)$; then $\pi_j = \lim\limits_{h \to +\infty} p_{ij}^{(h)}$ if and only if the chain is aperiodic.

To illustrate the results obtained in this section, consider a variant of the random walk Markov chain (2.8.7). We restrict the states to be the natural numbers $S = \{1, 2, 3, \cdots\}$ and define the transition matrix as:

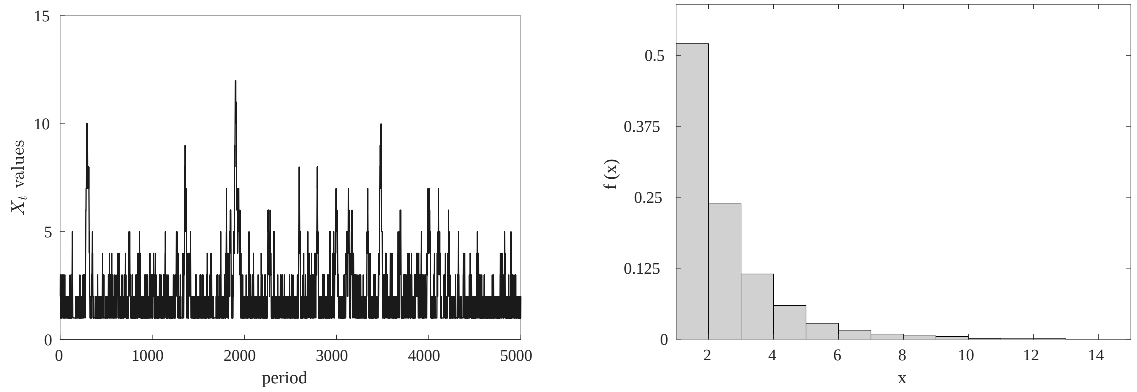
$$P = \begin{pmatrix} p+q & r & 0 & 0 & 0 & \cdots \\ p & q & r & 0 & 0 & \cdots \\ 0 & p & q & r & 0 & \cdots \\ 0 & 0 & p & q & r & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{2.8.10}$$

In state 1, the chain remains still with a probability of $p+q$, and moves up to 2 with a probability of r. In any other state, the chain remains still with a probability of q, moves up with a probability of r, and moves down with a probability of p. If it exists, the invariant distribution of the chain is given by (book 2, p. 25):

$$\pi_1 = 1 - \frac{r}{p} \quad , \quad \pi_2 = \left(\frac{r}{p}\right)\pi_1 \quad , \quad \pi_3 = \left(\frac{r}{p}\right)^2 \pi_1 \quad , \quad \pi_4 = \left(\frac{r}{p}\right)^3 \pi_1 \cdots \tag{2.8.11}$$

It is apparent from (2.8.11) that the invariant distribution exists if and only if $p > r$, in which case the probabilities π_j decline geometrically and all the states are positive recurrent from theorem 8.2. Also, the chain is clearly aperiodic so theorem 8.3 applies and $\pi_j = \lim\limits_{h \to +\infty} p_{ij}^{(h)}$: whatever the initial state of the chain, we converge to π_j for sufficiently large h.

So, assume we want to sample values from the invariant distribution (2.8.11), using algorithm 8.1. We set $p = 0.5$ and $q = r = 0.25$, which yields $\pi_1 = 0.5, \pi_2 = 0.25, \pi_3 = 0.125$, and so on. The chain is started at state 1 and run for 7000 periods, the first 2000 of which are discarded as burn-in sample. The simulated values and empirical distribution are displayed in Figure 8.4.



(a) Markov chain values X_t

(b) Empirical distribution of X

Figure 8.4: Distribution sampling with countable state Markov chain

The empirical distribution replicates the invariant distribution quite closely. The good fit is explained both by the large fraction of burn-in sample (2000 iterations) which permits convergence to the invariant distribution, and by the large number of iterations post transient sample, which from theorem 8.2.3 guarantees the convergence in mean to the true values.

8.3 Markov Chains with continuous state space

After finite and countable state spaces, we eventually discuss continuous state spaces. In this case, the Markov chain takes real values, and the set of possible states is $S = \mathbb{R}$ or some subset of it. Because the states are uncountable, it is not possible to define a transition matrix. Also, defining p_{ij} as the transition probability between states s_i and s_j is not sensitive anymore since the probability of any state is 0 on a continuous space.

Instead we use the notion of **transition density** or **transition kernel** $q(x,y)$. This is not the probability of moving from state y to state x. Rather, it represents the conditional density function $f(X_{t+1} = y|X_t = x)$. Then if the current state is $X_t = x$, the probability of moving to some subset A of S is given by:

$$\mathbb{P}(x,A) = \mathbb{P}(X_{t+1} \in A|X_t = x) = \int_A q(x,y)dy \tag{2.8.12}$$

The h-steps ahead transition kernel $q^{(h)}(x,y)$ is given by:

$$q^{(h)}(x,y) = \int_S q^{(h-1)}(x,z)\ q(z,y)\ dz \tag{2.8.13}$$

This says that in order to move from x to y after h periods, we first need to move from x to any state $z \in S$ in $h-1$ periods, then move from z to x over the last period. We then integrate over all possible intermediate states z to obtain the density $q^{(h)}(x,y)$. Following, we define the probability of moving to some subset A of S in h steps as $\mathbb{P}^{(h)}(x,A) = \int_A q^{(h)}(x,y)dy$.

The continuous state space analogue of the invariant distribution is given by the notion of invariant density.

definition 8.10: let X_t be a Markov chain with continuous state space S and transition kernel $q(x,y)$; an **invariant density** is a probability density function $\pi(y)$ which satisfies:

$$\pi(y) = \int_S \pi(x)q(x,y)dx$$

The notion of aperiodicty in continuous spaces is unchanged and still given by definition 8.6. Irreducibility on the other hand must be re-defined.

definition 8.11: let X_t be a Markov chain with continuous state space S and transition kernel $q(x,y)$, and let $\pi(x)$ be some density function on S; the chain is **π-irreducible** if for each subset A of S with $\pi(A) > 0$, there exists an h such that $\mathbb{P}^{(h)}(x,A) > 0$.

π-irreducibility is the continuous-space analogue of definitions 8.4 and 8.5 of irreducibility for countable state spaces. Finally, we need to define a continuous-space equivalent of the notion of recurrence.

definition 8.12: let X_t be a π-irreducible Markov chain.
The chain is **recurrent** if for each subset A of S with $\pi(A) > 0$:
$\mathbb{P}^{(h)}(x,A)\ i.o. > 0$ for all x
$\mathbb{P}^{(h)}(x,A)\ i.o. = 1$ for π-almost all x
The chain is **Harris recurrent** if $\mathbb{P}^{(h)}(x,A)\ i.o. = 1$ for all x

where $i.o$ stands for “infinitely often”. In short, the chain is recurrent if it returns to any subset A of S infinitely often with probability 1 for almost all initial states x. It is Harris recurrent if instead the condition holds for all x.

We then have the following theorem.

theorem 8.4: let X_t be a Markov chain with invariant distribution π, and suppose that X_t is π-irreducible. Then X_t is positive recurrent and π is the unique invariant distribution of X_t.
If X_t is also aperiodic, then for π-almost every x: $\left\| \mathbb{P}^{(h)}(x,A) - \pi(A) \right\| \to 0$,
with $\|.\|$ the total variation norm[1].
If X_t is Harris recurrent, then the convergence occurs for all x.

Theorem 8.4 constitutes the basis of modern Monte Carlo Markov Chain (MCMC) methods. It provides a simple procedure to sample from a target distribution. First, define a transition kernel that is irreducible, aperiodic, positive recurrent and whose invariant distribution corresponds to the target distribution.
Second, start the kernel from any state and run it for long enough to eventually sample values from the target distribution.

To illustrate the use of theorem 8.4, assume we want to sample values from a normal distribution with mean μ and variance σ: $\pi(y) \sim N(\mu,\sigma)$. To do so, we use an autoregressive transition kernel, defined as:

$$y_t = c + \gamma y_{t-1} + \varepsilon \qquad \varepsilon \sim N(0,s) \tag{2.8.14}$$

We claim that defining $c = \mu(1-\gamma)$ and $s = (1-\gamma^2)\sigma$, the unique invariant distribution of the transition kernel (2.8.14) is the target distribution $\pi(y) \sim N(\mu,\sigma)$. To see this, start from definition 8.10:

$$\begin{aligned} &\pi(y_{t-1})\ q(y_{t-1},y_t)\ dy_{t-1} \\ \propto\ &\int \exp\left(-\frac{1}{2}\frac{(y_{t-1}-\mu)^2}{\sigma}\right)\ \exp\left(-\frac{1}{2}\frac{(y_t - c - \gamma y_{t-1})^2}{s}\right) \end{aligned} \tag{2.8.15}$$

[1] The total variation norm between any two probability measures π_1 and π_2 is defined as:
$\|\pi_1(A) - \pi_2(A)\| = sup_A|\pi_1(A) - \pi_2(A)|$, for some set $A \in S$.

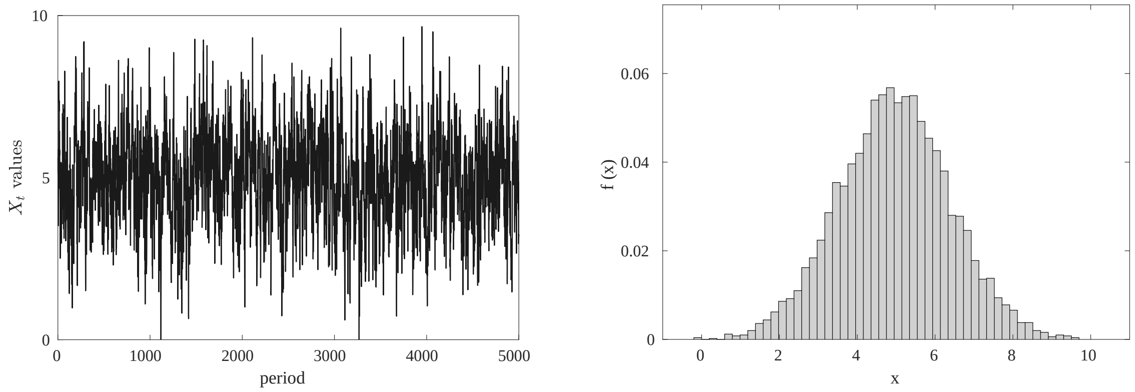
After some manipulations, this rewrites as (book 2, p. 26):

$$
\begin{aligned}
&= \exp\left(-\frac{1}{2}\frac{(y_t-\mu)^2}{\sigma}\right)\int \exp\left(-\frac{1}{2}\frac{(y_{t-1}-c-\gamma y_t)^2}{s}\right)\, dy_{t-1} \\
&\propto \exp\left(-\frac{1}{2}\frac{(y_t-\mu)^2}{\sigma}\right) \qquad (2.8.16)
\end{aligned}
$$

And this is indeed recognised as the density function of the target distribution $\pi(y_t)$.

The Markov chain defined by the transition kernel (2.8.14) is clearly π-irreducible, Harris recurrent and aperiodic. Thus from theorem 8.4 we known that it will converge to the invariant distribution π, provided it is run for a sufficient number of periods.

The target distribution is parameterized with $\mu = 5$ and $\sigma = 2$. The kernel uses $\gamma = 0.8$, and the chain is run for 3000 burn-in iterations and an additional 5000 draws. The simulations and the resulting empirical distribution are shown in Figure 8.5.



(a) Markov chain values X_t

(b) Empirical distribution of X

Figure 8.5: Distribution sampling with continuous state Markov chain

8.4 Application to Gibbs sampling

In this brief section, we demonstrate how the results obtained in the preceding sections justify the use of the Gibbs sampling algorithm. To keep the presentation simple, the analysis is restricted to the case of two parameters only, but the conclusions are general and extend to the case of n parameters.

Thus, consider a model with two parameters so that $\theta = \{\theta_1, \theta_2\}$. Our objective is to sample values from the posterior distribution $\pi(\theta|y) = \pi(\theta_1, \theta_2|y)$ which constitutes the target distribution. Marginalisation is not possible, but the conditional posteriors $\pi(\theta_1|y, \theta_2)$ and $\pi(\theta_2|y, \theta_1)$ are standard, so one can easily draw values from them. We use them to define a transition kernel $q(\theta^{(n-1)}, \theta^{(n)})$ that samples alternatively from both conditional posteriors:

$$
q(\theta^{(n-1)}, \theta^{(n)}) = \pi(\theta_1^{(n)}|\theta_2^{(n-1)})\ \pi(\theta_2^{(n)}|\theta_1^{(n)}) \qquad (2.8.17)
$$

We have dropped y in the conditionning for readibility. We now show that the target distribution $\pi(\theta_1, \theta_2)$ corresponds to the invariant distribution of the transition kernel (2.8.17).

Using definition 8.10, we obtain:

$$
\begin{aligned}
& \int q(\theta^{(n-1)},\theta^{(n)})\ \pi(\theta^{(n-1)})d\theta^{(n-1)} \\
= & \int \pi(\theta_1^{(n)}|\theta_2^{(n-1)})\ \pi(\theta_2^{(n)}|\theta_1^{(n)})\ \pi(\theta_1^{(n-1)},\theta_2^{(n-1)})d\theta_1^{(n-1)}d\theta_2^{(n-1)} \\
= & \ \pi(\theta_2^{(n)}|\theta_1^{(n)}) \int \pi(\theta_1^{(n)}|\theta_2^{(n-1)})\ \pi(\theta_2^{(n-1)})d\theta_2^{(n-1)} \\
= & \ \pi(\theta_2^{(n)}|\theta_1^{(n)})\ \pi(\theta_1^{(n)}) \\
= & \ \pi(\theta_1^{(n)},\theta_2^{(n)})
\end{aligned}
\tag{2.8.18}
$$

Hence, the target distribution $\pi(\theta_1^{(n)},\theta_2^{(n)})$ is the invariant distribution of the Gibbs sampler transition kernel $q(\theta^{(n-1)},\theta^{(n)}) = \pi(\theta_1^{(n)}|\theta_2^{(n-1)})\ \pi(\theta_2^{(n)}|\theta_1^{(n)})$. This represents a necessary but not sufficient condition to ensure the convergence of the kernel to the invariant distribution. Verifying that the conditions for convergence are satisfied may be difficult in general, but the following result establishes that the Gibbs sampling algorithm will work under mild assumptions.

theorem 8.5: let X_t be a Markov chain with invariant distribution π, and suppose that X_t is π-irreducible. If $\mathbb{P}(x,A)$ is absolutely continuous with respect to π for all x, then X_t is Harris recurrent.

The fact that the Gibbs sampling transition kernel can be Harris recurrent under mild conditions directly implies convergence to the invariant distribution, from theorem 8.4.

8.5 Application to Metropolis-Hastings

Unlike the Gibbs sampling algorithm, the Metropolis-Hastings algorithm does not require that we can sample values directly from the conditional posterior distributions. It uses a more general approach, built on the concept of **reversible kernel**. A transition kernel $q(x,y)$ is reversible if it satisfies:

$$\pi(x)\ q(x,y)\ =\ \pi(y)\ q(y,x) \tag{2.8.19}$$

We first show that if the transition kernel $q(x,y)$ is reversible, then $\pi(x)$ represents the invariant density for $q(x,y)$. From definition 8.10, we have:

$$\int \pi(x)q(x,y)dx = \int \pi(y)q(y,x)dx = \pi(y)\int q(y,x)dx = \pi(y) \tag{2.8.20}$$

So if we can find a reversible kernel $q(x,y)$, it becomes easy to sample values from the target distribution $\pi(x)$. In general however a transition kernel may not be reversible. In this case, we may obtain for instance that fome some x and y:

$$\pi(x)\ q(x,y)\ >\ \pi(y)\ q(y,x) \tag{2.8.21}$$

In this case, loosely speaking, the process moves from x to y too often, and from y to x too rarely. The trick consists in turning (2.8.21) into a reversible kernel by reducing the probability to move from x to y. To do so, we use a function $\alpha(x,y) < 1$ that represents the probability of move from x to y. If the move is not made, the process remains at x. With this, we obtain the reversible kernel:

$$\pi(x)\ q(x,y)\ \alpha(x,y) =\ \pi(y)\ q(y,x)\ \alpha(y,x) \tag{2.8.22}$$

Since the process moves from y to x too infrequently, the probability of move $\alpha(y,x)$ should be set to 1. But then this defines $\alpha(x,y)$ since (2.8.22) directly implies:

$$\alpha(x,y) = \frac{\pi(y)\ q(y,x)}{\pi(x)\ q(x,y)} \tag{2.8.23}$$

Conveniently, computing $\alpha(x,y)$ does not require the normalization constant of $\pi(.)$ since it appears both in the numerator and denominator.

For some values x and y the inequality (2.8.21) may be reversed so that $\pi(x)\ q(x,y)\ <\ \pi(y)\ q(y,x)$. In this case the process moves too rarely from x to y and we want $\alpha(x,y)$ to be 1 to compensate. Following, (2.8.23) becomes:

$$\alpha(x,y) = min\left\{1, \frac{\pi(y)\ q(y,x)}{\pi(x)\ q(x,y)}\right\} \tag{2.8.24}$$

Sampling from the target distribution $\pi(x)$ can then be done easily by defining a transition kernel $q(x,y)$ and adopting the probability of move (2.8.24). The conditions for converge specified in theorems 8.4 and 8.5 still apply.

PART III

Econometrics

CHAPTER 9

The linear regression model

This chapter introduces the most basic econometrics model: the linear regression. It focuses on its formulation and on the Bayesian estimates obtained under different assumptions. The associated applications (predictions and model selection) will be the object of chapter 10.

9.1 Formulation and maximum likelihood estimate

The linear regression model studies the relation between an **endogenous variable** y and a group of k **exogenous variables** $x_1, x_2, \cdots, x_k$ that explain it. To estimate the model, a sample of n observations is collected. The model then takes the form:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma) \qquad i = 1, \cdots, n \tag{3.9.1}$$

It is convenient to rewrite the model in compact form as:

$$y = X\beta + \varepsilon \qquad \varepsilon \sim N(0, \sigma I_n) \tag{3.9.2}$$

with:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{3.9.3}$$

The parameters of interest to estimate are then $\theta = \{\beta, \sigma\}$. Consider for now a frequentist approach of the model. Following section 3.1, we first need to set the likelihood function $f(y|\beta, \sigma)$ to obtain maximum likelihood estimates of β and σ. It follows immediately from equation (3.9.2) that $y \sim N(X\beta, \sigma I_n)$. The likelihood function is then given by:

$$f(y|\beta, \sigma) = (2\pi\sigma)^{-n/2} \exp\left(-\frac{1}{2}\frac{(y - X\beta)'(y - X\beta)}{\sigma}\right) \tag{3.9.4}$$

Following definition 3.5, the maximum likelihood estimates $\hat{\beta}$ and $\hat{\sigma}$ are obtained by maximizing the log-likelihood function:

$$\hat{\beta}, \hat{\sigma} = \underset{\beta, \sigma}{argmax} \ \log(f(y|\beta, \sigma)) \tag{3.9.5}$$

The log-likelihood function is given by:

$$\log(f(y|\beta, \sigma)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma) - \frac{1}{2}\frac{(y - X\beta)'(y - X\beta)}{\sigma} \tag{3.9.6}$$

The maximum is found by solving simultaneously for $\dfrac{\partial \log(f(y|\beta, \sigma))}{\partial \beta} = 0$ and $\dfrac{\partial \log(f(y|\beta, \sigma))}{\partial \sigma} = 0$.

It can be shown (book 2, p. 31) that the resulting estimates are:

$$\hat{\beta} = (X'X)^{-1}X'y \qquad \hat{\sigma} = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n} \tag{3.9.7}$$

The maximum likelihood estimate $\hat{\beta}$ is therefore equivalent to the standard OLS estimate. The estimate $\hat{\sigma}$ is also similar to the OLS estimate but is biased (the divisor is n instead of $n-k$).

A confidence interval at the α confidence level for any individual coefficient β_i can be obtained from (see for instance Greene (2003), chapter 4):

$$\hat{\beta}_i \pm T_{\alpha/2}\, s_i \qquad s_i = \sqrt{\hat{\sigma} S_{ii}} \qquad S = (X'X)^{-1} \qquad df = n-k \tag{3.9.8}$$

9.2 A first Bayesian estimate

The simplest Bayesian approach consists in treating σ as known so that only β remains to estimate. To do so, we define $\sigma = \hat{\sigma}$, the maximum likelihood estimate obtained in (3.9.7). In this case, we are left with $\theta = \{\beta\}$. From Bayes rule 3.3, the posterior $\pi(\beta|y)$ is given by:

$$\pi(\beta|y) \propto f(y|\beta)\pi(\beta) \tag{3.9.9}$$

The likelihood function $f(y|\beta)$ is given by (3.9.4). Consider then the prior distribution for β. Because the coefficients can take any real value, the multivariate normal distribution represents a good choice. We thus set the prior to be multivariate normal with prior mean b and prior variance V: $\pi(\beta) \sim N(b, V)$. Following:

$$\pi(\beta) = (2\pi)^{-k/2}|V|^{-1/2}\exp\left(-\frac{1}{2}(\beta - b)'V^{-1}(\beta - b)\right) \tag{3.9.10}$$

Following, Bayes rule (3.9.9) becomes:

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2}\frac{(y - X\beta)'(y - X\beta)}{\sigma}\right) \times \exp\left(-\frac{1}{2}(\beta - b)'V^{-1}(\beta - b)\right) \tag{3.9.11}$$

Notice the similarity between the linear regression and the stock return model developed in section 3.4: both models combine a normal likelihood function with a normal prior, the only difference being the multivariate nature of the regression. We basically follow the same estimation procedure, and in particular we apply again the "completing the squares" methodology. The details are provided once more due to the multivariate nature of the model, but they are essentially the same as in the scalar case. First develop and group the terms in (3.9.11) to obtain (book 2, p. 31):

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2}\left[\beta'(V^{-1} + \sigma^{-1}X'X)\beta - 2\beta'(V^{-1}b + \sigma^{-1}X'y) + b'V^{-1}b + y'\sigma^{-1}y\right]\right) \tag{3.9.12}$$

Now add terms in (3.9.12) to make the expression factorable into a single quadratic form.

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2}\left[\begin{array}{l} \beta'(V^{-1} + \sigma^{-1}X'X)\beta - 2\beta'\bar{V}^{-1}\bar{V}(V^{-1}b + \sigma^{-1}X'y) \\ +b'V^{-1}b + y'\sigma^{-1}y + \bar{b}'\bar{V}^{-1}\bar{b} - \bar{b}'\bar{V}^{-1}\bar{b} \end{array}\right]\right) \tag{3.9.13}$$

Define:

$$\bar{V} = (V^{-1} + \sigma^{-1}X'X)^{-1} \qquad \bar{b} = \bar{V}(V^{-1}b + \sigma^{-1}X'y) \tag{3.9.14}$$

Then (3.9.13) rewrites:

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2}(\beta'\bar{V}^{-1}\beta - 2\beta'\bar{V}^{-1}\bar{b} + \bar{b}'\bar{V}^{-1}\bar{b} + b'V^{-1}b + y'\sigma^{-1}y - \bar{b}'\bar{V}^{-1}\bar{b})\right) \tag{3.9.15}$$

Factoring the first three terms into a single quadratic form and separating the remaining terms yields:

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right) \times \exp\left(-\frac{1}{2}(b'V^{-1}b+y'\sigma^{-1}y-\bar{b}'\bar{V}^{-1}\bar{b})\right) \tag{3.9.16}$$

Noting that the second term does not involve β, relegate it to the normalization constant:

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right) \tag{3.9.17}$$

This is the kernel of a multivariate normal distribution with mean $\bar{b}$ and variance $\bar{V}$: $\pi(\beta|y) = N(\bar{b},\bar{V})$.

9.3 A hierarchical prior

This section considers a first model where both β and σ are estimated, so that $\theta = \{\beta,\sigma\}$. Following, Bayes rule is given by:

$$\pi(\beta,\sigma|y) \propto f(y|\beta,\sigma)\pi(\beta,\sigma) \tag{3.9.18}$$

We set a hierarchical prior by assuming that the prior distribution of β depends on the residual variance σ. Following, we have $\pi(\beta,\sigma) = \pi(\beta|\sigma)\pi(\sigma)$ and Bayes rule (3.9.18) rewrites:

$$\pi(\beta,\sigma|y) \propto f(y|\beta,\sigma)\pi(\beta|\sigma)\pi(\sigma) \tag{3.9.19}$$

The likelihood function $f(y|\beta,\sigma)$ for the model is still given by (3.9.4). For β, the hierarchical prior is set as a multivariate normal distribution with variance proportional to the residual variance σ: $\pi(\beta|\sigma) \sim N(b,\sigma V)$. Following:

$$\pi(\beta|\sigma) = (2\pi)^{-k/2}|\sigma V|^{-1/2}\exp\left(-\frac{1}{2}(\beta-b)'(\sigma V)^{-1}(\beta-b)\right) \tag{3.9.20}$$

For σ finally we use an inverse gamma prior with shape $\alpha/2$ and scale $\delta/2$: $\pi(\sigma) \sim IG(\alpha/2,\delta/2)$, so that:

$$\pi(\sigma) = \frac{\delta/2^{\alpha/2}}{\Gamma(\alpha/2)}\sigma^{-\alpha/2-1}\exp\left(-\frac{\delta}{2\sigma}\right) \tag{3.9.21}$$

Notice that this model is essentially the same as the hierarchical model developed in section 4.2 for the stock return example. We thus follow similar procedures,and obtain similar results. From Bayes rule (3.9.19), we obtain:

$$\begin{aligned}\pi(\beta,\sigma|y) \; &\propto \sigma^{-n/2}\exp\left(-\frac{1}{2}\frac{(y-X\beta)'(y-X\beta)}{\sigma}\right) \times |\sigma V|^{-1/2}\exp\left(-\frac{1}{2}(\beta-b)'(\sigma V)^{-1}(\beta-b)\right) \\ &\times \sigma^{-\alpha/2-1}\exp\left(-\frac{\delta}{2\sigma}\right)\end{aligned} \tag{3.9.22}$$

Grouping the terms and completing the squares, this joint posterior becomes (book 2, p. 32):

$$\pi(\beta,\sigma|y) \propto \sigma^{-k/2}\exp\left(-\frac{1}{2}(\beta-\bar{b})'(\sigma\bar{V})^{-1}(\beta-\bar{b})\right) \times \sigma^{-\bar{\alpha}/2-1}\exp\left(-\frac{\bar{\delta}}{2\sigma}\right) \tag{3.9.23}$$

with:

$$\bar{V} = (V^{-1}+X'X)^{-1} \qquad \bar{b} = \bar{V}(V^{-1}b+X'y) \qquad \bar{\alpha} = \alpha+n \qquad \bar{\delta} = \delta+y'y+b'V^{-1}b-\bar{b}'\bar{V}^{-1}\bar{b} \tag{3.9.24}$$

We recognize in the posterior the product of two kernels: a multivariate normal density, and an inverse gamma density. We are interested in the marginal posteriors $\pi(\beta|y)$ and $\pi(\sigma|y)$, and to do this we use definition 4.3. Marginalisation is easy for σ since β only appears in the first kernel, hence:

$$\begin{aligned}\pi(\sigma|y) = \int \pi(\beta,\sigma|y)d\beta \ &\propto \sigma^{-\bar{\alpha}/2-1} \exp\left(-\frac{\bar{\delta}}{2\sigma}\right) \int \sigma^{-k/2} \exp\left(-\frac{1}{2}(\beta-\bar{b})'(\sigma\bar{V})^{-1}(\beta-\bar{b})\right)d\beta \\ &\propto \sigma^{-\bar{\alpha}/2-1} \exp\left(-\frac{\bar{\delta}}{2\sigma}\right)\end{aligned} \tag{3.9.25}$$

This is the kernel of an inverse gamma distribution with shape $\bar{\alpha}/2$ and scale $\bar{\delta}/2$: $\pi(\sigma|y) \sim IG(\bar{\alpha},\bar{\delta})$.

The marginal posterior $\pi(\beta|y)$ is less direct. As σ appears in all the terms of (3.9.23), we group them and integrate:

$$\pi(\beta|y) = \int \pi(\beta,\sigma|y)d\sigma \ \propto \ \int \sigma^{-(\bar{\alpha}+k)/2-1} \exp\left(-\frac{\bar{\delta}+(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})}{2\sigma}\right)d\sigma \tag{3.9.26}$$

This is the kernel of an inverse Gamma distribution with shape $(\bar{\alpha}+k)/2$ and scale $(\bar{\delta}+(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})/2$, and integration yields the reciprocal of the normalization constant of the distribution. Hence:

$$\pi(\beta|y) \propto \Gamma\left(\frac{\bar{\alpha}+k}{2}\right)\left(\frac{\bar{\delta}+(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})}{2}\right)^{-\frac{\bar{\alpha}+k}{2}} \tag{3.9.27}$$

After some manipulations, it can be shown (book 2, p. 33) that this reformulates as:

$$\pi(\beta|y) \propto \left(1+\frac{1}{\bar{\alpha}}(\beta-\bar{b})'(\bar{\delta}\bar{V}/\bar{\alpha})^{-1}(\beta-\bar{b})\right)^{-\frac{\bar{\alpha}+k}{2}} \tag{3.9.28}$$

This is the kernel of a multivariate Student distribution with location $\bar{b}$, scale $\bar{\delta}\bar{V}/\bar{\alpha}$ and degrees of freedom $\bar{\alpha}$: $\pi(\beta|y) \sim T(\bar{b},\bar{\delta}\bar{V}/\bar{\alpha},\bar{\alpha})$.

9.4 An independent prior

This section introduces a second model where both β and σ are estimated. This time however β and σ are treated as independent parameters. Given that $\theta = \{\beta,\sigma\}$, Bayes rule is still given by (3.9.18). However, assuming independence yields $\pi(\beta,\sigma) = \pi(\beta)\ \pi(\sigma)$ so that:

$$\pi(\beta,\sigma|y) \propto f(y|\beta,\sigma)\pi(\beta)\pi(\sigma) \tag{3.9.29}$$

Using the likelihood function (3.9.4) and the priors (3.9.10) and (3.9.21), the joint posterior obtains as:

$$\begin{aligned}\pi(\beta,\sigma|y) \ &\propto \sigma^{-n/2} \exp\left(-\frac{1}{2}\frac{(y-X\beta)'(y-X\beta)}{\sigma}\right) \\ &\times \exp\left(-\frac{1}{2}(\beta-b)'V^{-1}(\beta-b)\right) \times \sigma^{-\alpha/2-1}\exp\left(-\frac{\delta}{2\sigma}\right)\end{aligned} \tag{3.9.30}$$

Again, any term not involving β or σ has been relegated to the normalization constant. Analytical marginalization from integration is not possible with this joint posterior. The situation is similar to the stock return example developed in section 6.1, and the solution also involves use of the Gibbs sampling algorithm. Obtain first the conditional posterior $\pi(\beta|y,\sigma)$. From definition 6.1, this is done by starting from the joint posterior (3.9.30) and relegating to the normalization constant any multiplicative term not involving β, yielding:

$$\pi(\beta|y,\sigma) \propto \exp\left(-\frac{1}{2}\frac{(y-X\beta)'(y-X\beta)}{\sigma}\right) \times \exp\left(-\frac{1}{2}(\beta-b)'V^{-1}(\beta-b)\right) \tag{3.9.31}$$

This is similar to (3.9.11), so rearranging and completing the squares the same way eventually results in:

$$\pi(\beta|y,\sigma) \propto \exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right) \tag{3.9.32}$$

with:

$$\bar{V} = (V^{-1}+\sigma^{-1}X'X)^{-1} \qquad \bar{b} = \bar{V}(V^{-1}b+\sigma^{-1}X'y) \tag{3.9.33}$$

This is the kernel of a multivariate normal distribution with mean $\bar{b}$ and variance $\bar{V}$: $\pi(\beta|y,\sigma) \sim N(\bar{b},\bar{V})$. Consider then the conditional posterior $\pi(\sigma|y,\beta)$. Start from (3.9.30), relegate to the normalization constant any multiplicative term not involving σ and rearrange to obtain:

$$\pi(\sigma|y,\beta) \propto \sigma^{-\bar{\alpha}/2-1}\ \exp\left(-\frac{\bar{\delta}}{2\sigma}\right) \tag{3.9.34}$$

with:

$$\bar{\alpha} = \alpha+n \qquad \bar{\delta} = \delta+(y-X\beta)'(y-X\beta) \tag{3.9.35}$$

This is the kernel of an inverse gamma distribution with shape $\bar{\alpha}/2$ and scale $\bar{\delta}/2$: $\pi(\sigma|y,\beta) \sim IG(\bar{\alpha}/2,\bar{\delta}/2)$.

We can now introduce the Gibbs sampling algorithm for the linear regression model.

algorithm 9.1: Gibbs sampling algorithm for the linear regression model

1. set initial values $\beta^{(0)}$ and $\sigma^{(0)}$. We use the maximum likelihood estimates $\beta^{(0)} = \hat{\beta}$ and $\sigma^{(0)} = \hat{\sigma}$.
2. at iteration j, draw:

   $\beta^{(j)}$ from $\pi(\beta|y,\sigma) \sim N(\bar{b},\bar{V})$ with:

   $\bar{V} = (V^{-1}+\sigma^{-1}X'X)^{-1} \qquad \bar{b} = \bar{V}(V^{-1}b+\sigma^{-1}X'y)$
3. at iteration j, draw:

   $\sigma^{(j)}$ from $\pi(\sigma|y,\beta) \sim IG(\bar{\alpha}/2,\bar{\delta}/2)$ with:

   $\bar{\alpha} = \alpha+n \qquad \bar{\delta} = \delta+(y-X\beta)'(y-X\beta)$
4. repeat until the desired number of iterations is realised.

9.5 Linear regression with heteroscedastic disturbances

The linear regression model assumes that the residual variance is constant over observations: $\varepsilon_i \sim N(0,\sigma)$. Sometimes this assumption is untenable and heteroscedasticity must be explicitely integrated in the model. The linear regression then reformulates as:

$$y_i = \beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\varepsilon_i \qquad \varepsilon_i \sim N(0,\sigma w_i) \qquad i=1,\cdots,n \tag{3.9.36}$$

The residual variance is now made observation-specific through the weighting term w_i. To estimate the model, we follow the approach of Koop (2003). First, assume that the weights w_i are a log-linear function of h regressors:

$$w_i = \exp(\gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_h z_{ih}) \qquad i = 1, \cdots, n \tag{3.9.37}$$

The h regressors $z_{i1}, \cdots, z_{ih}$ may include some or all of the regressors $x_{i1}, \cdots, x_{ik}$, and possibly other regressors. It may not include a constant, which would be redundant with the common variance term σ. For observation i the model rewrites in compact form as:

$$y_i = x_i'\beta + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma\ \exp(z_i'\gamma)) \qquad z_i = \begin{pmatrix} z_{i1} & z_{i2} & \cdots & z_{ih} \end{pmatrix}' \tag{3.9.38}$$

Stacking then for the n observations:

$$y = X\beta + \varepsilon \qquad \varepsilon \sim N(0, \sigma W) \qquad W = diag(\exp(Z\gamma)) \qquad Z = \begin{pmatrix} z_1 & z_2 & \cdots & z_n \end{pmatrix}' \tag{3.9.39}$$

The parameters of interest for the model are then $\theta = \{\beta, \sigma, \gamma\}$. Following definition 3.3, Bayes rule is given by:

$$\pi(\beta, \sigma, \gamma | y) \propto f(y|\beta, \sigma, \gamma)\pi(\beta, \sigma, \gamma) \tag{3.9.40}$$

From (3.9.39), the likelihood function obtains as (book 2, p. 33):

$$f(y|\beta, \sigma, \gamma) = (2\pi\sigma)^{-n/2}\ |W|^{-1/2}\ \exp\left(-\frac{1}{2}\frac{(y - X\beta)'W^{-1}(y - X\beta)}{\sigma}\right) \tag{3.9.41}$$

For further reference, it is useful to note that the likelihood function alternatively rewrites as:

$$f(y|\beta, \sigma, \gamma) = (2\pi\sigma)^{-n/2}\ \exp\left(-\frac{1}{2}\left[1_n'Z\gamma + (y - X\beta)'\ diag(\exp(-Z\gamma))\ (y - X\beta)/\sigma\right]\right) \tag{3.9.42}$$

For the prior we follow definition 4.1 and assume independence between the parameters so that $\pi(\beta, \sigma, \gamma) = \pi(\beta)\pi(\sigma)\pi(\gamma)$. The priors $\pi(\beta)$ and $\pi(\sigma)$ are respectively given by (3.9.10) and (3.9.21). For the prior $\pi(\gamma)$, we set a multivariate normal prior: $\pi(\gamma) \sim N(g, Q)$, so that:

$$\pi(\gamma) = (2\pi)^{-h/2}|Q|^{-1/2}\exp\left(-\frac{1}{2}(\gamma - g)'Q^{-1}(\gamma - g)\right) \tag{3.9.43}$$

Bayes rule (3.9.40) is not tractable analytically, so Gibbs sampling methods are required. Applying definition 6.1, the conditional posterior $\pi(\beta|y, \sigma, \gamma)$ obtains from the joint posterior (3.9.40) and relegating any term not involving β to the normalization constant. This yields $\pi(\beta|y, \sigma, \gamma) \propto f(y|\beta, \sigma, \gamma)\pi(\beta)$. Using the likelihood function (3.9.41) and the prior (3.9.10), one obtains:

$$\pi(\beta|y, \sigma, \gamma) \propto \exp\left(-\frac{1}{2}\frac{(y - X\beta)'W^{-1}(y - X\beta)}{\sigma}\right) \times \exp\left(-\frac{1}{2}(\beta - b)'V^{-1}(\beta - b)\right) \tag{3.9.44}$$

Completing the squares and rearranging yields (book 2, p. 34):

$$\pi(\beta|y, \sigma, \gamma) \propto \exp\left(-\frac{1}{2}(\beta - \bar{b})'\bar{V}^{-1}(\beta - \bar{b})\right) \tag{3.9.45}$$

with:

$$\bar{V} = (V^{-1} + \sigma^{-1}X'W^{-1}X)^{-1} \qquad \bar{b} = \bar{V}(V^{-1}b + \sigma^{-1}X'W^{-1}y) \tag{3.9.46}$$

This is the kernel of a multivariate normal distribution with mean $\bar{b}$ and variance $\bar{V}$: $\pi(\beta|y, \sigma, \gamma) = N(\bar{b}, \bar{V})$.

Consider now the conditional posterior $\pi(\sigma|y,\beta,\gamma)$. Start from the joint posterior (3.9.40) and relegate any term not involving σ to the normalization constant. This yields $\pi(\sigma|y,\beta,\gamma) \propto f(y|\beta,\sigma,\gamma)\pi(\sigma)$. Using the likelihood function (3.9.41) and the prior (3.9.21) then rearranging yields:

$$\pi(\sigma|y,\beta,\gamma) \propto \sigma^{-\bar{\alpha}/2-1}\ \exp\left(-\frac{\bar{\delta}}{2\sigma}\right) \tag{3.9.47}$$

with:

$$\bar{\alpha} = \alpha + n \qquad\qquad \bar{\delta} = \delta + (y - X\beta)'W^{-1}(y - X\beta) \tag{3.9.48}$$

Consider finally the conditional posterior $\pi(\gamma|y,\beta,\sigma)$. Start from the joint posterior (3.9.40) and relegate any term not involving γ to the normalization constant. This yields $\pi(\gamma|y,\beta,\sigma) \propto f(y|\beta,\sigma,\gamma)\pi(\gamma)$. Using the reformulated likelihood function (3.9.42) and the prior (3.9.43) yields:

$$\pi(\gamma|y,\beta,\sigma) \propto \exp\left(-\frac{1}{2}\left[1_n'Z\gamma + (y - X\beta)'\ diag(\exp(-Z\gamma))\ (y - X\beta)/\sigma + (\gamma - g)'Q^{-1}(\gamma - g)\right]\right) \tag{3.9.49}$$

This form is non-standard and cannot be rearranged into a known distribution. Sampling from the conditional posterior $\pi(\gamma|y,\beta,\sigma)$ thus requires the use of the Metropolis-Hastings algorithm. We choose a simple random walk kernel of the form:

$$\gamma^{(j)} = \gamma^{(j-1)} + e \qquad\qquad e \sim N(0, \tau I_h) \tag{3.9.50}$$

This implies that $q(\gamma^{(j-1)}, \gamma^{(j)}) \sim N(\gamma^{(j-1)}, \tau I_h)$, with τ an exogenous hyperparameter set to generate a 20-30% acceptance rate of the algorithm. Using definition 7.2 and noting that the symmetry of the kernel implies $q(\gamma^{(j-1)}, \gamma^{(j)}) = q(\gamma^{(j)}, \gamma^{(j-1)})$, the acceptance probability is given by $\alpha(\gamma^{(j-1)}, \gamma^{(j)}) = min\{1, \pi(\gamma^{(j)}|y,\beta,\sigma)/\pi(\gamma^{(j-1)}|y,\beta,\sigma)\}$. Given (3.9.49), this yields:

$$\begin{aligned} &\alpha(\gamma^{(j-1)}, \gamma^{(j)}) \\ =\ & \min\left\{1, \exp\left(-\frac{1}{2}\left[\begin{array}{l} 1_n'Z(\gamma^{(j)} - \gamma^{(j-1)}) \\ +(y - X\beta)'\ diag[\exp(-Z\gamma^{(j)}) - \exp(-Z\gamma^{(j-1)})]\ (y - X\beta)/\sigma \\ +(\gamma^{(j)} - g)'Q^{-1}(\gamma^{(j)} - g) - (\gamma^{(j-1)} - g)'Q^{-1}(\gamma^{(j-1)} - g) \end{array}\right]\right)\right\} \end{aligned} \tag{3.9.51}$$

The Gibbs sampling algorithm for the model with heteroscedasticity is then:

algorithm 9.2: Gibbs sampling algorithm for the linear regression model with heteroscedasticity

1. set initial values $\beta^{(0)}$, $\sigma^{(0)}$ and $\gamma^{(0)}$. We use the maximum likelihood estimates $\beta^{(0)} = \hat{\beta}$ and $\sigma^{(0)} = \hat{\sigma}$, and set $\gamma^{(0)} = 0$.

2. at iteration j, draw:

   $\beta^{(j)}$ from $\pi(\beta|y,\sigma,\gamma) \sim N(\bar{b}, \bar{V})$ with:

   $\bar{V} = (V^{-1} + \sigma^{-1}X'W^{-1}X)^{-1} \qquad \bar{b} = \bar{V}(V^{-1}b + \sigma^{-1}X'W^{-1}y)$

3. at iteration j, draw:

   $\sigma^{(j)}$ from $\pi(\sigma|y,\beta,\gamma) \sim IG(\bar{\alpha}/2, \bar{\delta}/2)$ with:

   $\bar{\alpha} = \alpha + n \qquad \bar{\delta} = \delta + (y - X\beta)'W^{-1}(y - X\beta)$

4. at iteration j, draw:

   a candidate value $\tilde{\gamma}$ from $\tilde{\gamma} = \gamma^{(j-1)} + e$, $\pi(e) \sim N(0, \tau I_h)$

5. at iteration j: obtain the acceptance probability $\alpha(\gamma^{(j-1)},\gamma^{(j)})$ given by (3.9.51)
6. at iteration j: draw a uniform random number u from $u \sim U(0,1)$.

   if $u \leq \alpha(\gamma^{(j-1)},\gamma^{(j)})$, set $\gamma^{(j)} = \tilde{\gamma}$; else, set $\gamma^{(j)} = \gamma^{(j-1)}$
7. repeat until the desired number of iterations is realised.

9.6 Linear regression with autocorrelated disturbances

Consider the linear regression model in the context of time series. It is common in this case that the disturbances display serial correlation across periods or autocorrelation. The model may then rewrite as:

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t \qquad \varepsilon_t = \phi_1 \varepsilon_{t-1} + \cdots + \phi_q \varepsilon_{t-q} + u_t \qquad u_t \sim N(0,\sigma) \tag{3.9.52}$$

The sample contains T observations for $t = 1,\cdots,T$, and at each period the disturbance ε_t is related to q of its lags (autocorrelation of order q). The model rewrites in compact form as:

$$y_t = x_t'\beta + \varepsilon_t \qquad \varepsilon_t = e_t'\phi + u_t \qquad u_t \sim N(0,\sigma) \tag{3.9.53}$$

with:

$$x_t = \begin{pmatrix} x_{1t} & x_{2t} & \cdots & x_{kt} \end{pmatrix}' \qquad e_t = \begin{pmatrix} \varepsilon_{t-1} & \varepsilon_{t-2} & \cdots & \varepsilon_{t-q} \end{pmatrix}' \qquad \phi = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_q \end{pmatrix}' \tag{3.9.54}$$

The parameters of interest of the model are then $\theta = \{\beta,\sigma,\phi\}$. To estimate the model, we follow the approach of Chib (1993).

From definition 3.3 and assuming independence between the parameters as in definition 4.1 so that $\pi(\beta,\sigma,\phi) = \pi(\beta)\pi(\sigma)\pi(\phi)$, Bayes rule is given by:

$$\pi(\beta,\sigma,\phi|y) \propto f(y|\beta,\sigma,\phi)\pi(\beta)\pi(\sigma)\pi(\phi) \tag{3.9.55}$$

Consider first the likelihood function $f(y|\beta,\sigma,\phi)$. For the incoming developements, we define the **lag polynomial** $\phi(L)$ as:

$$\phi(L)x_t \;=\; (1-\phi_1 L - \cdots - \phi_q L^q)x_t \;=\; x_t - \phi_1 x_{t-1} - \cdots - \phi_q x_{t-q} \qquad L^r x_t \equiv x_{t-r} \tag{3.9.56}$$

Apply the lag polynomial on both sides of (3.9.53) and rewrite in compact form for the T periods to obtain:

$$y^* = X^*\beta + u \qquad u \sim N(0,\sigma I_T) \tag{3.9.57}$$

with:

$$y^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_T^* \end{pmatrix} \qquad y_t^* \equiv \phi(L)y_t \qquad X^* = \begin{pmatrix} x_1^{*\prime} \\ x_2^{*\prime} \\ \vdots \\ x_T^{*\prime} \end{pmatrix} \qquad x_t^* \equiv \phi(L)x_t \qquad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{pmatrix} \tag{3.9.58}$$

We assume that q initial conditions are available to compute y_t^* and x_t^* for $t = 1,2,\cdots$ It follows immediately from (3.9.57) that $y^* \sim N(X^*\beta,\sigma I_T)$. The likelihood function then writes as:

$$f(y|\beta,\sigma,\phi) = (2\pi\sigma)^{-T/2} \exp\left(-\frac{1}{2}\frac{(y^*-X^*\beta)'(y^*-X^*\beta)}{\sigma}\right) \tag{3.9.59}$$

Alternatively, rewrite (3.9.53) as:

$$\varepsilon = E\phi + u \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix} \qquad E = \begin{pmatrix} e_1' \\ e_2' \\ \vdots \\ e_T' \end{pmatrix} \tag{3.9.60}$$

It then follows that $\varepsilon \sim N(E\phi, \sigma I_T)$ and the likelihood function rewrites as:

$$f(y|\beta,\sigma,\phi) = (2\pi\sigma)^{-T/2} \exp\left(-\frac{1}{2}\frac{(\varepsilon - E\phi)'(\varepsilon - E\phi)}{\sigma}\right) \tag{3.9.61}$$

For the priors, $\pi(\beta)$ and $\pi(\sigma)$ are unchanged and respectively given by (3.9.10) and (3.9.21). For ϕ, we assume a multivariate normal distribution with mean p and variance H: $\pi(\phi) \sim N(p,H)$. Following:

$$\pi(\phi) = (2\pi)^{-q/2}|H|^{-1/2}\exp\left(-\frac{1}{2}(\phi - p)'H^{-1}(\phi - p)\right) \tag{3.9.62}$$

Bayes rule (3.9.55) is not tractable analytically, so Gibbs sampling methods are required. Applying definition 6.1, the conditional posterior $\pi(\beta|y,\sigma,\phi)$ obtains from the joint posterior (3.9.55) and relegating any term not involving β to the normalization constant. This yields $\pi(\beta|y,\sigma,\phi) \propto f(y|\beta,\sigma,\phi)\pi(\beta)$. Using the likelihood function (3.9.59) and the prior (3.9.10), one obtains:

$$\pi(\beta|y,\sigma,\phi) \propto \exp\left(-\frac{1}{2}\frac{(y^* - X^*\beta)'(y^* - X^*\beta)}{\sigma}\right) \times \exp\left(-\frac{1}{2}(\beta - b)'V^{-1}(\beta - b)\right) \tag{3.9.63}$$

This is similar to (3.9.11) (with y^* and X^* instead of y and X), so after completing the squares, we obtain:

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2}(\beta - \bar{b})'\bar{V}^{-1}(\beta - \bar{b})\right) \tag{3.9.64}$$

with:

$$\bar{V} = (V^{-1} + \sigma^{-1}X^{*\prime}X^*)^{-1} \qquad \bar{b} = \bar{V}(V^{-1}b + \sigma^{-1}X^{*\prime}y^*) \tag{3.9.65}$$

This is the kernel of a multivariate normal distribution with mean $\bar{b}$ and variance $\bar{V}$: $\pi(\beta|y,\sigma,\phi) = N(\bar{b},\bar{V})$.

Consider now the conditional posterior $\pi(\sigma|y,\beta,\phi)$. Start from the joint posterior (3.9.55) and relegate any term not involving σ to the normalization constant. This yields $\pi(\sigma|y,\beta,\phi) \propto f(y|\beta,\sigma,\phi)\pi(\sigma)$. Using the likelihood function (3.9.59) and the prior (3.9.21) then rearranging yields:

$$\pi(\sigma|y,\beta,\phi) \propto \sigma^{-\bar{\alpha}/2-1}\exp\left(-\frac{\bar{\delta}}{2\sigma}\right) \tag{3.9.66}$$

with:

$$\bar{\alpha} = \alpha + T \qquad \bar{\delta} = \delta + (y^* - X^*\beta)'(y^* - X^*\beta) \tag{3.9.67}$$

This is the kernel of an inverse gamma distribution with shape $\bar{\alpha}$ and scale $\bar{\delta}$: $\pi(\sigma|y,\beta,\phi) \sim IG(\bar{\alpha},\bar{\delta})$.

Consider finally the conditional posterior $\pi(\phi|y,\beta,\sigma)$. Start from the joint posterior (3.9.55) and relegate any term not involving ϕ to the normalization constant. This yields $\pi(\phi|y,\beta,\sigma) \propto f(y|\beta,\sigma,\phi)\pi(\phi)$. Using the reformulated likelihood function (3.9.61) and the prior (3.9.62) then rearranging yields:

$$\pi(\phi|y,\beta,\sigma) \propto \exp\left(-\frac{1}{2}\frac{(\varepsilon - E\phi)'(\varepsilon - E\phi)}{\sigma}\right) \times \exp\left(-\frac{1}{2}(\phi - p)'H^{-1}(\phi - p)\right) \tag{3.9.68}$$

Completing the squares and rearranging yields (book 2, p. 34):

$$\pi(\phi|y,\beta,\sigma) \propto \exp\left(-\frac{1}{2}(\phi-\bar{p})'\bar{H}^{-1}(\phi-\bar{p})\right) \tag{3.9.69}$$

with:

$$\bar{H}=(H^{-1}+\sigma^{-1}E'E)^{-1} \qquad \bar{p}=\bar{H}(H^{-1}p+\sigma^{-1}E'\varepsilon) \tag{3.9.70}$$

This is the kernel of a multivariate normal distribution with mean $\bar{p}$ and variance $\bar{H}$: $\pi(\phi|y,\beta,\sigma) \sim N(\bar{p},\bar{H})$.

The Gibbs sampling algorithm for the model with autocorrelation is then:

algorithm 9.3: Gibbs sampling algorithm for the linear regression model with autocorrelation

1. set initial values $\beta^{(0)}$, $\sigma^{(0)}$ and $\phi^{(0)}$. We use the maximum likelihood estimates $\beta^{(0)}=\hat{\beta}$, $\sigma^{(0)}=\hat{\sigma}$ and set $\phi^{(0)}=0$.
2. at iteration j, draw:
   $\beta^{(j)}$ from $\pi(\beta|y,\sigma,\phi) \sim N(\bar{b},\bar{V})$ with:
   $(V^{-1}+\sigma^{-1}X^{*\prime}X^{*})^{-1} \qquad \bar{b}=\bar{V}(V^{-1}b+\sigma^{-1}X^{*\prime}y^{*})$
3. at iteration j, draw:
   $\sigma^{(j)}$ from $\pi(\sigma|y,\beta,\phi) \sim IG(\bar{\alpha}/2,\bar{\delta}/2)$ with:
   $\bar{\alpha}=\alpha+T \qquad \bar{\delta}=\delta+(y^{*}-X^{*}\beta)'(y^{*}-X^{*}\beta)$
4. at iteration j, draw:
   $\phi^{(j)}$ from $\pi(\phi|y,\beta,\sigma) \sim N(\bar{p},\bar{H})$ with:
   $\bar{H}=(H^{-1}+\sigma^{-1}E'E)^{-1} \qquad \bar{p}=\bar{H}(H^{-1}p+\sigma^{-1}E'\varepsilon)$
5. repeat until the desired number of iterations is realised.

9.7 Efficient estimation

Consider the Bayesian regression model with independent prior developed in section 9.4. The model necessitates the Gibbs sampling algorithm and thus implies that at each iteration a new value β is sampled from its conditional posterior $\pi(\beta|y,\sigma) \sim N(\bar{b},\bar{V})$ (see step 2 in algorithm 9.1). The parameters for the posterior are given by (3.9.33), repeated here for convenience:

$$\bar{V}=(V^{-1}+\sigma^{-1}X'X)^{-1} \qquad \bar{b}=\bar{V}(V^{-1}b+\sigma^{-1}X'y) \tag{3.9.71}$$

Notice that the computation of $\bar{b}$ involves the calculation of $\bar{V}$, and that the computation of $\bar{V}$ in turn implies an explicit matrix inversion. Inversion is a costly operation, with the cost increasing at a cubic rate with k, the dimension of $\bar{V}$. For small values of k inversion can be performed quickly, but for large values the cost may become prohibitive, especially since the calculation must be repeated at each iteration of the Gibbs algorithm.

To save some computational time, it is preferable to adopt an alternative approach which avoids the explicit inversion required to compute $\bar{V}$. First, note from (3.9.71) that we can calculate $\bar{V}^{-1}=(V^{-1}+\sigma^{-1}X'X)$ without inversion of the right-hand side. Then denote by G the lower triangular Cholesky factor of $\bar{V}^{-1}$ so that $\bar{V}^{-1}=GG'$. This in turn implies that $\bar{V}=(GG')^{-1}=G^{-1\prime}G^{-1}$ so that $G^{-1\prime}$ is the (upper triangular) Cholesky factor of $\bar{V}$.

Note then that from property d.2 of the multivariate normal distribution we can sample a value β from $\pi(\beta|y,\sigma) \sim N(\bar{b},\bar{V})$ by calculating:

$$\beta = \bar{b} + \xi \qquad\qquad \xi \sim N(0,\bar{V}) \tag{3.9.72}$$

And equivalently, this can be done from:

$$\beta = G^{-1\prime}G^{-1}(V^{-1}b + \sigma^{-1}X'y) + G^{-1\prime}\zeta \qquad\qquad \zeta \sim N(0,I_k) \tag{3.9.73}$$

Eventually factoring the $G^{-1\prime}$ term yields:

$$\beta = G^{-1\prime}\left[G^{-1}(V^{-1}b + \sigma^{-1}X'y) + \zeta\right] \qquad\qquad \zeta \sim N(0,I_k) \tag{3.9.74}$$

Sampling a value β then only involves an inversion of G, twice. The benefit of (3.9.74) is that G is a triangular matrix so that inversion can be done at a cheaper cost by back- and forward-substitution. Such optimized inversion procedures for triangular matrices are routinely performed by numerical softwares. It can then be shown (see for instance Golub and Loan (1996)) that using this approach is twice as fast as using brute strength inversion in (3.9.71), which proves critical for large dimensional models[1].

The method is general and can be summarised by the following algorithm:

algorithm 9.4: Efficient sampling algorithm

Consider some n-dimensional parameter θ with $\theta \sim N(\mu,\Sigma)$ where Σ^{-1} is known, μ is of the form $\mu = \Sigma m$, and m is some known n-dimensional vector. To sample efficiently from $\theta \sim N(\mu,\Sigma)$:

1. compute G, the Cholesky factor of Σ^{-1}, so that $\Sigma^{-1} = GG'$.
2. sample ζ from $\zeta \sim N(0,I_n)$.
3. solve for $\theta = G^{-1\prime}\left[G^{-1}m + \zeta\right]$ efficiently by back- and forward-substitution.

Algorithm 9.4 can be applied to any model involving a normal distribution and an explicit inversion of its variance matrix. In particular, it can also be used to reduce the computational cost of the β steps in algorithms 9.2 and 9.3 for the heteroscedastic and autocorrelated regression models.

9.8 Application: estimating a Taylor rule for the United States

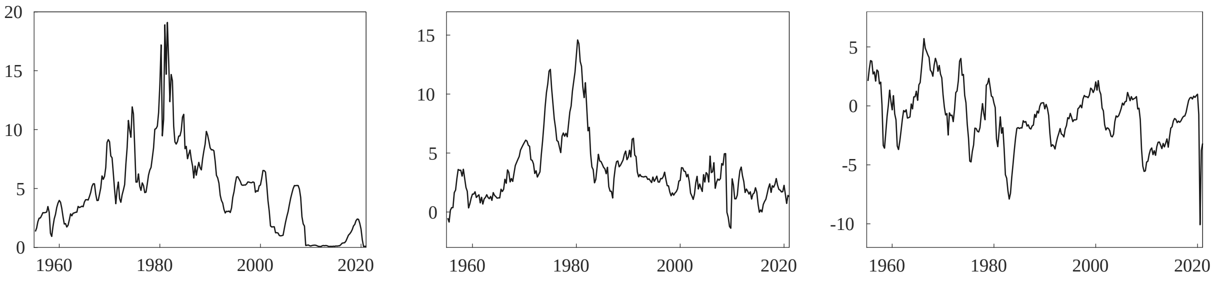
The conduct of monetary policy constitutes the core activity of central banking institutions. To understand how central institutions determine the leading interest rate, Taylor (1993) proposed a simple targetting rule linking the nominal interest rate to inflation and the output gap. Precisely, he postulated that central banks respond to inflation and economic activity with a linear policy rule of the form:

$$r = \bar{r} + \gamma\pi + \phi\hat{y} \tag{3.9.75}$$

r denotes the federal funds rate, $\bar{r}$ the target real interest rate, while π and $\hat{y}$ respectively denote the inflation rate and output gap, defined as the percentage deviation of actual output from potential output. γ and ϕ are the policy parameters determining the amplitude of the response of central authorities. For the policy parameters, Taylor (1993) assumed values of $\bar{r} = 1$, $\gamma = 1.5$ and $\phi = 0.5$. This implies that the FED responds to positive inflation and output gap with contractionary monetary policy, increasing the federal funds rate in reaction to inflationary and overheating pressures.

[1] Precisely, the number of operations to invert a generic $k \times k$ matrix is of order $2\mathcal{O}(k^3/3)$, while matrix inversion with Gauss-Jordan elimination only scales to $\mathcal{O}(k^3/3)$.

To verify the relevance of the Taylor rule for the US, we collect quarterly data for the federal funds rate, inflation and the output gap. The data is quarterly and ranges from 1955q1 to 2020q4[2]. The series are plotted in Figure 9.1.



(a) Federal funds rate (b) Inflation (c) Output gap

Figure 9.1: Time series for the US Taylor rule

We first start with a naive maximum likelihood estimate of the model. Table 9.1 reports the estimates of the Taylor rule for the different models developed in this chapter. Looking at the second line, we find our naive maximum likelihood estimate. The values are overall consistent with the theoretical Taylor rule, but the constant $\bar{r}$ is too large while the policy responses γ and ϕ are considerably too low. Also, the coefficients somewhat suffer from large standard errors, especially the constant $\bar{r}$.

model	$\bar{r}$	γ	ϕ
Taylor rule	1	1.5	0.5
maximum likelihood	1.36 [0.23]	0.99 [0.05]	0.14 [0.06]
simple Bayesian	1.03 [0.09]	1.11 [0.03]	0.34 [0.04]
hierarchical	1.17 [0.16]	1.04 [0.04]	0.21 [0.05]
independent	1.02 [0.09]	1.11 [0.03]	0.35 [0.04]
heteroscedastic	1.02 [0.08]	1.11 [0.04]	0.35 [0.03]
autocorrelated	1.02 [0.10]	0.88 [0.07]	0.38 [0.04]

Table 9.1: Posterior estimates for the US Taylor rule (standard deviations in square brackets)

We now try to improve the estimates with Bayesian methods. The five Bayesian models introduced in this chapter require the definition of a prior distribution $\pi(\beta) \sim N(b,V)$ for the regression coefficients. For the prior mean b, we can simply use the values implied by the theoretical Taylor rule. For the prior variance V, the choice becomes quite subjective. A reasonable strategy consists in setting V as a diagonal matrix, implying no a priori covariance between the regression coefficients. Also, for the variances, the following is proposed: assume that with 95% confidence the target rate $\bar{r}$ is comprised between 0.8 and 1.2. This implies a standard deviation of 0.1 and a variance of 0.01. Similarly, assuming with 95% confidence that the response to inflation γ is comprised between 1.3 and 1.7 yields a standard deviation of 0.1 and a variance of 0.01. Finally, if we believe with 95% confidence that the response to the output gap ϕ lies between 0.4 and 0.6, we obtain a standard deviation of 0.05 and a prior variance of 0.0025.

[2] The three series are obtained from the Saint Louis FED website; federal funds rate: series FEDFUNDS; inflation: series CPIAUCSL, switched to year-on-year growth rate; output gap: series GDPC1 of actual GDP, and GDPPOT of potential GDP; output gap defined as 100 times the ratio of actual over potential GDP.

This eventually yields:

$$b = \begin{pmatrix} 1 \\ 1.5 \\ 0.5 \end{pmatrix} \qquad V = \begin{pmatrix} 0.01 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 0.0025 \end{pmatrix} \tag{3.9.76}$$

Depending on the models, other priors have to be defined. For $\pi(\sigma) \sim IG(\alpha/2, \delta/2)$, a classical choice consists in setting $\alpha = \delta = 0.0001$. These tiny values set a diffuse prior, leaving the burden of estimation to the data. Similarly for the heteroscedatic and autocorrelated models, there are no obvious a priori values for $\pi(\gamma) \sim N(g, Q)$ and $\pi(\phi) \sim N(p, Z)$. So we set diffuse prior by setting $g = 0_h$, $Q = 100I_h$, $p = 0_q$, and $Z = 100I_q$.

The resulting estimates (using the posterior median) for the Bayesian models are displayed in rows 3-7 of Table 9.1. Two main conclusions arise. First, compared to the maximum likelihood regression, the Bayesian estimates get closer to the theoretical Taylor rule. The Bayesian priors effectively managed to mitigate the data information, driving the estimates towards the prior values. The obtained posterior estimates are thus more consistent with economic theory. Second, the addition of prior information also contributed to reduce the posterior variance, producing more accurate estimates. This is especially obvious for the constant $\bar{r}$, but overall all the coefficients benefited from the additional prior insight.

The question that arises next is: are these Bayesian models really better than the regular maximum likelihood model? Do they produce better predictions? And among them, which one is the most relevant? These questions will be answered in the next chapter.

CHAPTER 10

Applications with the linear regression model

This chapter introduces two essential features of the linear regression model: prediction, and model selection.

10.1 Prediction

Prediction is probably the most important application when it comes to the linear regression model. In the context of a frequentist approach with maximum likelihood estimates for β and σ, prediction is straightforward. Denote by $\hat{X}$ the $m \times k$ matrix containing the m additional vectors of regressors from which we want to predict, and by $\hat{y}$ the resulting m-dimensional vector of predictions. From (3.9.2), a minimum variance linear prediction obtains as:

$$\hat{y} = \mathbb{E}(y|\hat{X}) = \hat{X}\hat{\beta} \tag{3.10.1}$$

Confidence intervals at the α confidence level can then be obtained from (see for instance Greene (2003), chapter 6):

$$\hat{y} \pm T_{\alpha/2}(\sigma I_m + \hat{X}[\sigma(X'X)^{-1}]\hat{X}') \qquad df = n-k \tag{3.10.2}$$

In a Bayesian context predictions are formed using the posterior predictive distribution. Consider first the simple Bayesian regression developed in section 9.2. From definition 4.8, the likelihood function (3.9.4) and the posterior distribution (3.9.17), the posterior predictive distribution obtains as:

$$\begin{aligned} f(\hat{y}|y) &= \int f(\hat{y}|y,\beta)\ \pi(\beta|y)\ d\beta \\ &\propto \int \exp\left(-\frac{1}{2}\frac{(\hat{y}-\hat{X}\beta)'(\hat{y}-\hat{X}\beta)}{\sigma}\right)\ \exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right)\ d\beta \end{aligned} \tag{3.10.3}$$

After some algebraic manipulations, this rewrites as (book 2, p. 37):

$$f(\hat{y}|y) \propto \exp\left(-\frac{1}{2}(\hat{y}-\hat{X}\bar{b})'(\sigma I_m + \hat{X}\bar{V}\hat{X}')^{-1}(\hat{y}-\hat{X}\bar{b})\right) \tag{3.10.4}$$

where $\bar{b}$ and $\bar{V}$ are defined as in (3.9.14). This is the kernel of a multivariate normal distribution with mean $\hat{X}\bar{b}$ and variance $\sigma I_m + \hat{X}\bar{V}\hat{X}'$: $f(\hat{y}|y) \sim N(\hat{X}\bar{b},\ \sigma I_m + \hat{X}\bar{V}\hat{X}')$. The prediction is thus normal, centered on the posterior mean $\hat{X}\bar{b}$. The variance is similar to the scale parameter in the frequentist equation (3.10.2), except that the Bayesian estimate $\bar{V} = (V^{-1} + \sigma^{-1}X'X)^{-1}$ additionally accounts for the prior variance V.

Notice that the structure of the variance implies that the prediction has two sources of variance. The first component σI_m is the variance due to the intrinsic noise in the model (the residual term ε in (3.9.2)). The second component $\hat{X}\bar{V}\hat{X}'$ reflects the uncertainty about β, the unknown parameter of the model.

We now consider the regression model with the hierarchical prior developed in section 9.3. From definition 4.8, the likelihood function (3.9.4) and the posterior distribution (3.9.22), the posterior predictive distributon obtains as:

$$\begin{aligned} f(\hat{y}|y) &= \int\int f(\hat{y}|y,\beta,\sigma)\ \pi(\beta,\sigma|y)\ d\beta d\sigma \\ &\propto \int\int \sigma^{-m/2}\exp\left(-\frac{1}{2}\frac{(\hat{y}-\hat{X}\beta)'(\hat{y}-\hat{X}\beta)}{\sigma}\right)\ \exp\left(-\frac{1}{2}\frac{(y-X\beta)'(y-X\beta)}{\sigma}\right) \\ &\times\sigma^{-k/2}\exp\left(-\frac{1}{2}(\beta-b)'(\sigma V)^{-1}(\beta-b)\right)\times\sigma^{-\alpha/2-1}\exp\left(-\frac{\delta}{2\sigma}\right) \end{aligned} \tag{3.10.5}$$

After some manipulations, the expression reformulates as (book 2, p. 39):

$$f(\hat{y}|y)\propto\left(1+\frac{1}{\bar{\alpha}}(\hat{y}-\hat{X}\bar{b})'[\bar{\delta}(I_m+\hat{X}\bar{V}\hat{X}')/\bar{\alpha}]^{-1}(\hat{y}-\hat{X}\bar{b})\right)^{-(\bar{\alpha}+m)/2} \tag{3.10.6}$$

with $\bar{V},\bar{b},\bar{\alpha}$ and $\bar{\delta}$ defined as in (3.9.24). This is the kernel of a multivariate Student distribution with location $\hat{X}\bar{b}$, scale $\bar{\delta}(I_m+\hat{X}\bar{V}\hat{X}')/\bar{\alpha}$ and degrees of freedom $\bar{\alpha}$: $f(\hat{y}|y)\sim T(\hat{X}\bar{b},\bar{\delta}(I_m+\hat{X}\bar{V}\hat{X}')/\bar{\alpha},\bar{\alpha})$. Notice the similarities with (3.10.4): the predictive distribution is the same as in the Gaussian case, except that treating σ as an unknown parameter results in additional uncertainty. Following, the predictive distribution becomes Student, the fat tails reflecting the increased variance.

Consider predictions for the independent prior model developed in section 9.4. The model requires Gibbs sampling for estimation, and thus the predictive density must be recovered from the Gibbs sampling sampling draws, following algorithm 6.3. Adapted to the independent prior linear regression, the algorithm becomes:

algorithm 10.1: Gibbs sampling algorithm for the posterior predictive distribution, linear regression with independent prior

1. at iteration j, draw $\beta^{(j)}$ and $\sigma^{(j)}$ from their posterior distributions. Recycle the values obtained from the j^{th} iteration of the Gibbs sampling algorithm.

2. draw ε from $\varepsilon\sim N(0,\sigma I_m)$, then calculate $\hat{y}=\hat{X}\beta+\varepsilon$.

3. marginalize, that is, discard β and σ and keep only $\hat{y}$.

4. repeat until the desired number of iterations is realised.

Predictions are only slightly more complicated to obtain in the case of the heteroscedastic model of section 9.5 and the autocorrelation model of section 9.6. For the former, algorithm 6.3 becomes:

algorithm 10.2: Gibbs sampling algorithm for the posterior predictive distribution, linear regression with heteroscedasticity

1. at iteration j, draw $\beta^{(j)}$, $\sigma^{(j)}$ and $\gamma^{(j)}$ from their posterior distributions. Recycle the values obtained from the j^{th} iteration of the Gibbs sampling algorithm.

2. calculate W from $W=diag(\exp(\hat{Z}\gamma))$, then draw ε from $\varepsilon\sim N(0,\sigma W)$; finally, calculate $\hat{y}=\hat{X}\beta+\varepsilon$.

3. marginalize, that is, discard β,σ and γ and keep only $\hat{y}$.

4. repeat until the desired number of iterations is realised.

For the model with autocorrelation, finally, algorithm 6.3 becomes:

algorithm 10.3: Gibbs sampling algorithm for the posterior predictive distribution, linear regression with autocorrelation

1. at iteration j, draw $\beta^{(j)}$, $\sigma^{(j)}$ and $\phi^{(j)}$ from their posterior distributions. Recycle the values obtained from the j^{th} iteration of the Gibbs sampling algorithm.
2. for $j = 1, \cdots, m$, draw u_{t+j} from $u_{t+j} \sim N(0, \sigma)$, then calculate: $\varepsilon_{t+j} = \phi_1 \varepsilon_{t+j-1} + \cdots + \phi_q \varepsilon_{t+j-q} + u_{t+j}$.
3. for $j = 1, \cdots, m$, calculate $\hat{y}_{t+j}$ from $\hat{y}_{t+j} = x'_{t+j}\beta + \varepsilon_{t+j}$
4. marginalize, that is, discard β, σ and ϕ and keep only $\hat{y} = \hat{y}_{t+1}, \cdots, \hat{y}_{t+m}$.
5. repeat until the desired number of iterations is realised.

10.2 Forecast evaluation

Producing accurate predictions constitutes a central concern in linear gression. In this respect, forecast evaluation criteria constitutes an important aspect of the prediction exercise. We start the analysis with simple measures of in-sample fit. It follows immediately from (3.9.2) that $\varepsilon = y - X\beta$. Denoting by $\hat{\beta}$ the point estimate for the regression coefficients (the posterior median for all Bayesian models), an estimate of the residuals obtains as:

$$\hat{\varepsilon} = y - X\hat{\beta} \tag{3.10.7}$$

Based upon this, we define the following classical goodness of fit quantities: the sum of squared residuals, the R^2 and the adjusted-R^2:

$$SSR = \hat{\varepsilon}'\hat{\varepsilon} \qquad TSS = (y - \bar{y})'(y - \bar{y}) \qquad R^2 = 1 - \frac{SSR}{TSS} \qquad \text{adj-R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{3.10.8}$$

For the maximum likelihood regression, two additional classical measures of goodness of fit are provided by the Akaike Information Criterion (AIC) and the so-called Schwarz's Bayesian Information Criterion (BIC), respectively defined as:

$$AIC = 2|\theta|/n - 2\,\hat{L}/n \qquad BIC = |\theta|\log(n)/n - 2\,\hat{L}/n \tag{3.10.9}$$

where $|\theta|$ denote the number of parameters in the models, and $\hat{L} = \log(f(y|\hat{\theta}))$ is the log-likelihood of the model defined in (3.9.6), evaluated at the maximum likelihood estimates $\hat{\beta}$ and $\hat{\sigma}$. After some manipulations (book 2, p. 42), the two criteria rewrite as:

$$AIC = 2\,(k+1)/n + \log(2\pi) + \log(\hat{\varepsilon}'\hat{\varepsilon}/n) + 1 \qquad BIC = (k+1)\,\log(n)/n + \log(2\pi) + \log(\hat{\varepsilon}'\hat{\varepsilon}/n) + 1 \tag{3.10.10}$$

While in-sample criteria provide useful insight, most often we are interested in the out-of-sample predictive performance of the model. Denote again by $\hat{y}$ the m-dimensional vector of out-of-sample predictions, and by $\hat{y}_i$ the individual predictions in the vector, $i = 1, \cdots, m$. For Bayesian models, the predicted value is simply defined as the median of the posterior predictive distribution. Denote then by y_i, $i = 1, \cdots, m$ the set of corresponding actual values. Then the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are defined as:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \qquad MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \qquad MAPE = \frac{100}{m}\sum_{i=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.10.11}$$

The three criteria represent a measure of distance so that the lower the better. Two additional measures are of interest. The Theil inequality coefficient (Theil-U) is always comprised between 0 and 1; a perfect fit yields a value of 0, while forecasts get increasingly inaccurate as the value approaches 1. The bias represents the tendency of the prediction to be systematically higher of systematically lower than the realized value; A value of 0 indicates no bias, while values tending towards 1 (respectively -1) represents a tendency to be systematically higher (respectively lower) than the observation. The formulas are given by:

$$\text{Theil-U} = \frac{\sqrt{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{m} y_i^2}+\sqrt{\sum_{i=1}^{m}\hat{y}_i^2}} \qquad \text{bias} = \frac{\sum_{i=1}^{m} y_i-\hat{y}_i}{\sum_{i=1}^{m}|y_i-\hat{y}_i|} \tag{3.10.12}$$

The above forecast evaluation criterai consider only single-valued forecasts. Bayesian models however are richer since they result in full predictive distributions. Ideally, a Bayesian criterion should thus account for the entire distribution, and not just the point estimate. Intuitively, a preditive distribution will produce good forecasts if the realised values are located in points of high density. The log score (LogS) of Good (1952) and the continuous ranked probability score (CRPS) of Matheson and Winkler (1976) build on this principle. They are defined as:

$$LogS = -\ \log(\hat{f}(y_i)) \qquad CRPS = \int_{-\infty}^{+\infty}\left[\hat{F}(z)-\mathbb{1}(y_i\leq z)\right]^2 dz \tag{3.10.13}$$

where we use $\hat{f}$ and $\hat{F}$ to denote respectively the density and cumulative distribution functions of the predictive density. In short, both measures attribute a penalty to predictions that deviate from the points of high density in the predictive distribution. More accurate forecasts result in lower penalties and hence lower scores. Computing the log score is straightforward for the simple and hierarchical linear regressions since the predictive density take analytical forms (refer to (3.10.4) and (3.10.6), respectively). For the other models, one must rely on numerical approximations. A classical solution proposed by Krüger et al. (2017) consists in using a Gaussian approximation of the posterior predictive distribution, noting that predictive distributions are typically close to a Normal distribution. In this case, the log score is given by:

$$LogS = -\ \log(\hat{\phi}(y_i)) \tag{3.10.14}$$

where $\hat{\phi}$ denotes the density function of the normal distribution with mean $\hat{\mu}$ and variance $\hat{\sigma}$ calculated from the Gibbs sampler draws of the empirical predictive density.

For the CRPS, (3.10.13) is never used directly. Analytical equivalents are available whenever the predictive density is normal (Gneiting et al. (2005)) or Student (Jordan et al. (2019)). Consider first a Normal predictive density $f(\hat{y}_i|y)\sim N(\hat{\mu}_i,\hat{\sigma}_i)$, and denote by $\tilde{y}_i=(y_i-\hat{\mu}_i)/\hat{s}_i$, where $\hat{s}_i=\sqrt{\hat{\sigma}_i}$ is the standard deviation of the predictive distribution. Then the CRPS is given by:

$$CRPS = \hat{s}_i\left\{\ \tilde{y}_i(2\Phi(\tilde{y}_i)-1)+2\phi(\tilde{y}_i)-\frac{1}{\sqrt{\pi}}\ \right\} \tag{3.10.15}$$

where ϕ and Φ respectively denote the density and cumulative distribution function of the standard normal distribution. Consider then a Student distribution predictive density $f(\hat{y}_i|y)\sim T(\hat{\mu}_i,\hat{\sigma}_i,\hat{v}_i)$, and denote by $\tilde{y}_i=(y_i-\hat{\mu}_i)/\hat{s}_i$, where $\hat{s}_i=\sqrt{\hat{\sigma}_i}$ is the square root of the scale parameter. Then the formula becomes:

$$CRPS = \hat{s}_i\left\{\ \tilde{y}_i(2F(\tilde{y}_i)-1)+2f(\tilde{y}_i)\left(\frac{\hat{v}_i+\tilde{y}_i^2}{\hat{v}_i-1}\right)-\frac{2\sqrt{\hat{v}_i}}{\hat{v}_i-1}\ \frac{B(\frac{1}{2},\hat{v}_i-\frac{1}{2})}{B(\frac{1}{2},\frac{\hat{v}_i}{2})^2}\ \right\} \tag{3.10.16}$$

where $B(x)$ denotes the Beta function. Whenever analytical formulas are not available for the predictive density, the CRPS can be approximated from the Gibbs sampling draws of the posterior predictive distribution. Krüger et al. (2017) show that the CRPS can be consistently estimated from:

$$CRPS \approx \frac{1}{J}\sum_{j=1}^{J}|\hat{y}^{(j)}-y_i| - \frac{1}{2\,J^2}\sum_{j=1}^{J}\sum_{k=1}^{J}|\hat{y}^{(j)}-\hat{y}^{(k)}| \tag{3.10.17}$$

where $\hat{y}^{(j)}$ denotes draw j obtained from the Gibbs sampler for the predictive distribution. Equations (3.10.14) - (3.10.17) provide formulas for individual forecasts. For an exercise involving m forecasts, the overall log score and CRPS are then obtained by taking the mean over the m individual values.

10.3 Marginal likelihood

The marginal likelihood constitutes the basis of model comparison and hypothesis testing in the context of linea regression. Consider first the simple Bayesian regression developed in section 9.2. From definition 4.6, the marginal likelihood obtains from:

$$\begin{aligned} f(y) &= \int f(y|\beta)\pi(\beta)d\beta \\ &= \int (2\pi\sigma)^{-n/2}\ \exp\left(-\frac{1}{2}\frac{(y-X\beta)'(y-X\beta)}{\sigma}\right)\times(2\pi)^{-k/2}|V|^{-1/2}\exp\left(-\frac{1}{2}(\beta-b)'V^{-1}(\beta-b)\right)d\beta \end{aligned} \tag{3.10.18}$$

where use has been made of the likelihood function $f(y|\beta)$ given by (3.9.4) and the prior $\pi(\beta)$ given by (3.9.10). Rearranging and completing the squares, this reformulates as (book 2, p. 43):

$$\begin{aligned} f(y) &= (2\pi)^{-n/2}\ \sigma^{-n/2}|\bar{V}|^{1/2}|V|^{-1/2}\times\exp\left(-\frac{1}{2}\left[y'\sigma^{-1}y+b'V^{-1}b-\bar{b}\bar{V}^{-1}\bar{b}\right]\right) \\ &\times\int(2\pi)^{-k/2}|\bar{V}|^{-1/2}\exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right)d\beta \end{aligned} \tag{3.10.19}$$

with $\bar{V}$ and $\bar{b}$ defined as in (3.9.14). The second term is the probability density function of a multivariate normal distribution which thus integrates to 1, leaving only:

$$f(y) = (2\pi)^{-n/2}\ \sigma^{-n/2}|\bar{V}|^{1/2}|V|^{-1/2}\times\exp\left(-\frac{1}{2}\left[y'\sigma^{-1}y+b'V^{-1}b-\bar{b}\bar{V}^{-1}\bar{b}\right]\right) \tag{3.10.20}$$

Numerical instability may occur if the prior variance values in V are small. For this reason, it is convenient to reformulate (3.10.20) in numerically stable form as (book 2, p. 44):

$$f(y) = (2\pi)^{-n/2}\ \sigma^{-n/2}|I_k+\sigma^{-1}VX'X|^{-1/2}\exp\left(-\frac{1}{2}\left[y'\sigma^{-1}y+b'V^{-1}b-\bar{b}\bar{V}^{-1}\bar{b}\right]\right) \tag{3.10.21}$$

Next, consider the hierarchical prior developed in section 9.3. From definition 4.6, the marginal likelihood obtains from:

$$\begin{aligned} f(y) &= \int\int f(y|\beta,\sigma)\pi(\beta,\sigma)d\beta d\sigma \\ &= \int\int(2\pi\sigma)^{-n/2}\ \exp\left(-\frac{1}{2}\frac{(y-X\beta)'(y-X\beta)}{\sigma}\right) \\ &\times(2\pi)^{-k/2}|\sigma V|^{-1/2}\exp\left(-\frac{1}{2}(\beta-b)'(\sigma V)^{-1}(\beta-b)\right)\times\frac{\delta/2^{\alpha/2}}{\Gamma(\alpha/2)}\sigma^{-\alpha/2-1}\exp\left(-\frac{\delta}{2\sigma}\right)d\beta d\sigma \end{aligned} \tag{3.10.22}$$

where we used the likelihood function $f(y|\beta,\sigma)$ given by (3.9.4) and the priors $\pi(\beta|\sigma)$ given by (3.9.20) and $\pi(\sigma)$ given by (3.9.21). Rearranging and completing the squares, this reformulates as (book 2, p. 44):

$$f(y) = \pi^{-n/2}\, |V|^{-1/2} |\bar{V}|^{1/2}\, \frac{\delta^{\alpha/2}}{\bar{\delta}^{\bar{\alpha}/2}}\, \frac{\Gamma(\bar{\alpha}/2)}{\Gamma(\alpha/2)}$$
$$\times \int\int (2\pi)^{-k/2} |\sigma\bar{V}|^{-1/2} \exp\left(-\frac{1}{2}(\beta-\bar{b})'(\sigma\bar{V})^{-1}(\beta-\bar{b})\right) \times \frac{\bar{\delta}/2^{\bar{\alpha}/2}}{\Gamma(\bar{\alpha}/2)} \sigma^{-\bar{\alpha}/2-1} \exp\left(-\frac{\bar{\delta}}{2\sigma}\right) d\beta d\sigma \tag{3.10.23}$$

The values of $\bar{V}$, $\bar{b}$, $\bar{\alpha}$ and $\bar{\delta}$ are as in (3.9.24). The terms within the integrals respectively represent the probability density functions of multivariate normal and inverse Gamma distributions. They thus integrate to 1, leaving only:

$$f(y) = \pi^{-n/2}\, |V|^{-1/2} |\bar{V}|^{1/2}\, \frac{\delta^{\alpha/2}}{\bar{\delta}^{\bar{\alpha}/2}}\, \frac{\Gamma(\bar{\alpha}/2)}{\Gamma(\alpha/2)} \tag{3.10.24}$$

It is also convenient to reformulate this term in numerically stable form as (book 2, p. 46):

$$f(y) = \pi^{-n/2}\, |I_k + VX'X|^{-1/2}\, \frac{\delta^{\alpha/2}}{\bar{\delta}^{\bar{\alpha}/2}}\, \frac{\Gamma(\bar{\alpha}/2)}{\Gamma(\alpha/2)} \tag{3.10.25}$$

Consider then the independent prior developed in section 9.4. The model relies on simulation methods, so the marginal likelihood must be computed from equation (2.6.15), namely:

$$f(y) \approx \frac{f(y|\beta^*,\sigma^*)\pi(\beta^*,\sigma^*)}{\pi(\sigma^*|y,\beta^*) \times \frac{1}{J}\sum_{j=1}^{J} \pi(\beta^*|\sigma^{(j)},y)} \tag{3.10.26}$$

Using the likelihood function (3.9.4), the priors (3.9.10) and (3.9.21), and the conditional posteriors (3.9.33) and (3.9.35), it can be shown that the marginal likelihood formulates as (book 2, p. 47):

$$f(y) \approx \pi^{-n/2} \frac{\exp\left(-\frac{1}{2}(\beta-b)'V^{-1}(\beta-b)\right)}{\frac{1}{J}\sum_{j=1}^{J} |I_k + \sigma^{-1}VX'X)|^{1/2} \exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right)}\, \frac{\delta^{\alpha/2}}{\bar{\delta}^{\bar{\alpha}/2}}\, \frac{\Gamma(\bar{\alpha}/2)}{\Gamma(\alpha/2)} \tag{3.10.27}$$

This form is similar to (3.10.25), save for the approximation of the determinant term stemming from the Gibbs sampler.

Finally, we consider the linear regression models with heteroscedasticity and autocorrelation developed in sections 9.5 and 9.6, respectively. These models involve 3 blocks of parameters, and the heteroscedastic regression additionally necessitates the Metropolis-Hastings algorithm. For these reasons, the Chib (1995) approach cannot be used directly, and instead we use the Gelfand and Dey (1994) methodology introduced in section 7.4.

For the heteroscedastic model, a direct application of (2.7.17) yields:

$$\frac{1}{f(y)} \approx \frac{1}{J}\sum_{j=1}^{J} \frac{g(\theta^{(j)})}{f(y|\beta^{(j)},\sigma^{(j)},\gamma^{(j)})\; \pi(\beta^{(j)})\; \pi(\sigma^{(j)})\; \pi(\gamma^{(j)})} \tag{3.10.28}$$

Using the probability density function (2.7.19) along with the likelihood function (3.9.41) and the priors (3.9.10), (3.9.21) and (3.9.43) then rearranging yields (book 2, p. 48):

$$\log(f(y)) \approx -log\left((\omega J)^{-1}(2\pi)^{(n-1)/2}\, |\hat{\Sigma}|^{-1/2}\, |V|^{1/2}\, |Q|^{1/2}\, \frac{\Gamma(\alpha/2)}{\delta/2^{\alpha/2}}\right)$$
$$-log\left(\sum_{j=1}^{J} \mathbb{1}(\theta \in \hat{\Theta})\, |W|^{1/2}\, \sigma^{(\alpha+n)/2+1} \exp\left(\frac{1}{2}\begin{bmatrix} (y-X\beta)'(\sigma W)^{-1}(y-X\beta) + (\beta-b)'V^{-1}(\beta-b) \\ +\delta\sigma^{-1} + (\gamma-g)'Q^{-1}(\gamma-g) - (\theta-\hat{\theta})'\hat{\Sigma}^{-1}(\theta-\hat{\theta}) \end{bmatrix}\right)\right) \tag{3.10.29}$$

The summation term may easily break down when n gets large due to the $\sigma^{(\alpha+n)/2+1}$ term. A numerically stable solution consists in converting the log-summation into a summation of logs, which can be done using the so-called log-sum-exp identity:

$$\log\left(\sum_{j=1}^{J} x_i\right) = \log(\bar{x}) + \log\left(\sum_{j=1}^{J} \exp\Big(\log(x_i) - \log(\bar{x})\Big)\right) \qquad \bar{x} = max\{x_i\} \tag{3.10.30}$$

A similar strategy is applied to the regression model with autocorrelation: applying (2.7.17), we obtain:

$$\frac{1}{f(y)} \approx \frac{1}{J}\sum_{j=1}^{J} \frac{g(\theta^{(j)})}{f(y|\beta^{(j)},\sigma^{(j)},\gamma^{(j)})\ \pi(\beta^{(j)})\ \pi(\sigma^{(j)})\ \pi(\phi^{(j)})} \tag{3.10.31}$$

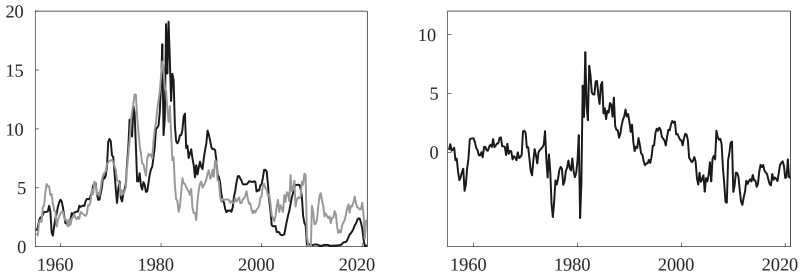
Using the probability density function (2.7.19) along with the likelihood function (3.9.61) and the priors (3.9.10), (3.9.21) and (3.9.62) then rearranging yields (book 2, p. 49):

$$\begin{aligned}\log(f(y)) \approx & -\log\left((\omega J)^{-1}(2\pi)^{(T-1)/2}\ |\hat{\Sigma}|^{-1/2}\ |V|^{1/2}\ |Z|^{1/2}\ \frac{\Gamma(\alpha/2)}{\delta/2^{\alpha/2}}\right) \\ & -\log\left(\sum_{j=1}^{J} \mathbb{1}(\theta \in \hat{\Theta})\sigma^{(\alpha+T)/2+1} \exp\left(\frac{1}{2}\left[\begin{array}{l}(\varepsilon - E\phi)'\sigma^{-1}(\varepsilon - E\phi) + (\beta - b)'V^{-1}(\beta - b) \\ +\delta\sigma^{-1} + (\phi - p)'Z^{-1}(\phi - p) - (\theta - \hat{\theta})'\hat{\Sigma}^{-1}(\theta - \hat{\theta})\end{array}\right]\right)\right)\end{aligned} \tag{3.10.32}$$

10.4 Application: revisiting the US Taylor rule

We return to the Taylor rule example developed in section 9.8, and ask two additional questions. Does the data exhibit heteroscedasticity or autocorrelation? And which model produces the best predictions?

To get a hint on the first question, we start by plotting the residuals obtained from the naive maximum likelihood regression (Figure 10.1).



(a) Actual and fitted

(b) Residuals

Figure 10.1: Maximum likelihood regression: fitted and residuals

At first sight, the residuals appear both heteroscedastic and autocorrelated. Clearly, their variance is not constant, especially in the 1980's which exhibits a fueling in volatility. Also, the residuals reveal periods of positive (1980-2000) and negative (1970-1980) autocorrelation. To make this point formal, we rely on the marginal likelihood setting developed in section 4.7. We compare the marginal likelihood of the independent Bayesian regression model based on the assumption of spherical disturbances with the marginal likelihood of the heteroscedastic and autocorrelated models. The values $m(y)$ of the different Bayesian regression models are reported in Table 10.1.

Model	*SSR*	adj-R^2	$m(y)$	*RMSE*	*LogS*	*CRPS*
max. likelihood	1398	0.589	–	3.19	–	–
simple Bayesian	1496	0.561	-268.38	2.11	2.16	1.22
hierarchical	1412	0.586	-267.48	2.48	2.33	1.45
independent	1503	0.559	-273.16	2.09	2.18	1.21
heteroscedastic	1481	0.565	-256.86	2.10	2.14	1.22
autocorrelated	1725	0.493	-199.74	1.79	2.12	1.06

Table 10.1: Forecast evaluation criteria and marginal likelihood for the linear regression models

The results are unambiguous: the model with spherical disturbances yields a marginal likelihood of -273.16, considerably smaller than the heteroscedastic model (-256.86) and the autocorrelated model (-199.74). Using Jeffrey's Guidelines provided in Table 4.1, we conclude that the data rejects the null hypothesis of spherical disturbances in favor of both heteroscedasticity and autocorrelation, the latter being most strongly supported.

We now consider the question of the best predictor. To do so, we separate the data sample into a train sample including the first 75% of the data (until 2004), and keep the remaining data as test sample. The models are first estimated on the train sample, along with in-sample fit scores (*SSR* and adjusted-R^2). Predictions are then formed on the test sample, and the forecasts are evaluated from prediction criteria (RMSE, log scores and CRPS).

By construction the maximum likelihood model obtains the best in-sample scores, closely followed by the hierarchical model. The simple, independent and heteroscedastic models perform average, while the autocorrelated looks especially poor.

Those in-sample results may be quite misleading though, and indeed the conclusions change radically whenever the models are considered for out-of-sample predictions. Two conclusions arise. First, the naive maximum likelihood proves the worst model in terms of forecast performance. It is beaten by every single Bayesian model in terms of RMSE, and quite significantly. This shows that adding relevant prior information does contribute to improve the predictive performance, while on the other hand simple OLS models tend to overfit.

Second, the autocorrelated models proves by far the best predictor. This is spectacular in terms of RMSE and CRPS, and remains marginally true for the log score. This comes in contrast with the poor in-sample performance, confirming that the quality of a model comes primarily from its ability to catch the true data generative process outside of the training sample. The heteroscedastic model also performs fair, but only to a lesser extent compared to the other Bayesian models. Overall, these results confirm the previous conclusion that heteroscedasticity and autocorrelation represent the correct behaviour of the data.

Bibliography

Chib, S. (1993). Bayes regression with autoregressive errors: A gibbs sampling approach. *Journal of Econometrics*, 58(3):275–294.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313 – 1321.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96:270–281.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269 – 281.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56.

Geweke, J. (1999). Using simulation methods for bayesian econometric models: inference, development,and communication. *Econometric Reviews*, 18:1–73.

Gneiting, T., Westveld, A. H., Raftery, A. E., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. Technical report, Monthly Weather Review.

Golub, G. H. and Loan, C. F. V. (1996). *Matrix Computations*. The Johns Hopkins University Press, 3^{rd} edition.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14:107–114.

Greenberg, E. (2012). *Introduction to Bayesian Econometrics*. Cambridge University Press, 2 edition.

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education, 5 edition.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press, Oxford, 3^{th} edition.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoring rules. *Journal of Statistical Software*, 90:1–37.

Koop, G. M. (2003). *Bayesian econometrics*. John Wiley & Sons Inc.

Krüger, F., Lerch, S., Thorarinsdottir, T. L., and Gneiting, T. (2017). Probabilistic forecasting and comparative model assessment based on markov chain monte carlo output. Technical report.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22:1087–1096.

Poirier, D. (1995). *Intermediate statistics and econometrics: a comparative approach*. The MIT Press.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy*, 39:195–214.

Subject index

Author index