Bayesian Time-Series Econometrics

Book 1 - theory

Romain Legrand

Second 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: $P(A) = \lim_{n \to \infty} \frac{m}{n}$

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 4$^{\text{th}}$ 2000 is not a repeatable random experiment since July 4$^{\text{th}}$ 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 $\mathbb{P}(B)$).

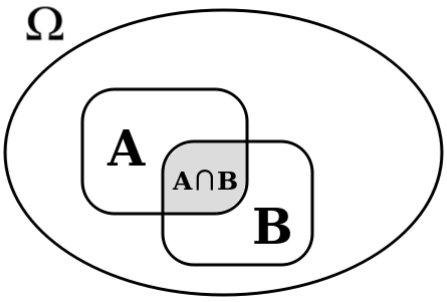


Figure 2.1: A representation of conditional probabilities with an Euler diagram

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, $f(1) + f(0) = 1/2 + 1/2 = 1$.

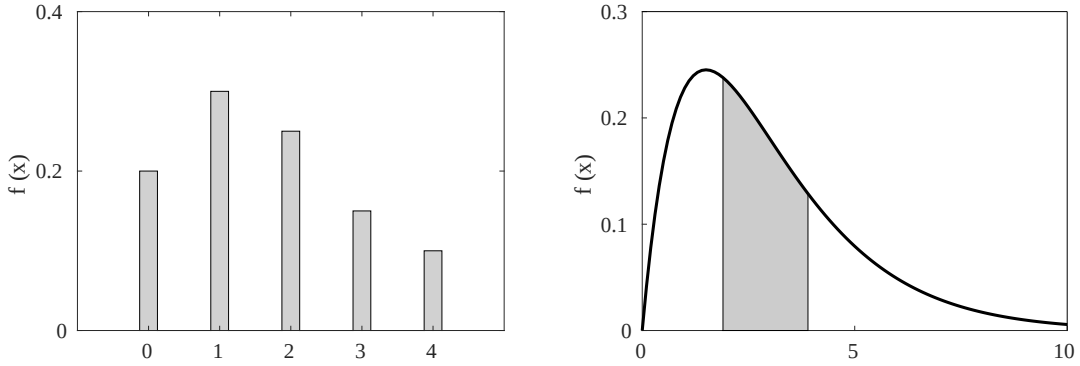
By contrast, continuous random take an uncountable number of possible values so that the probability of obtaining any single value is 0. Probabilities then only make sense over continuous intervals, which yields the notion of probability density function.

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:

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

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

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

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.

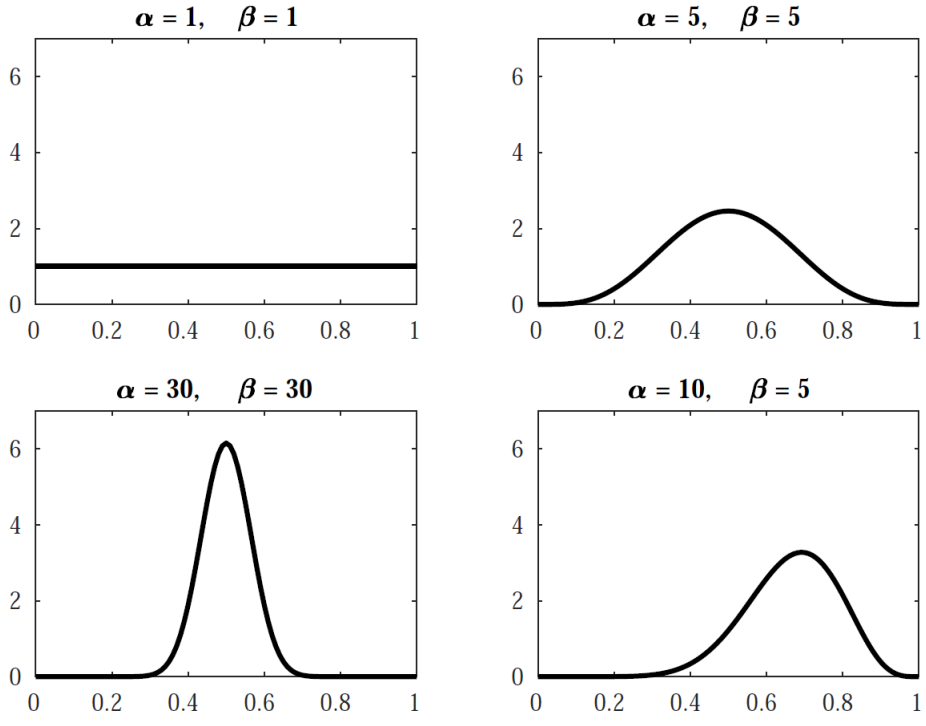


Figure 3.1: Probability density function of the Beta distribution for different α and β values

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.27 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.

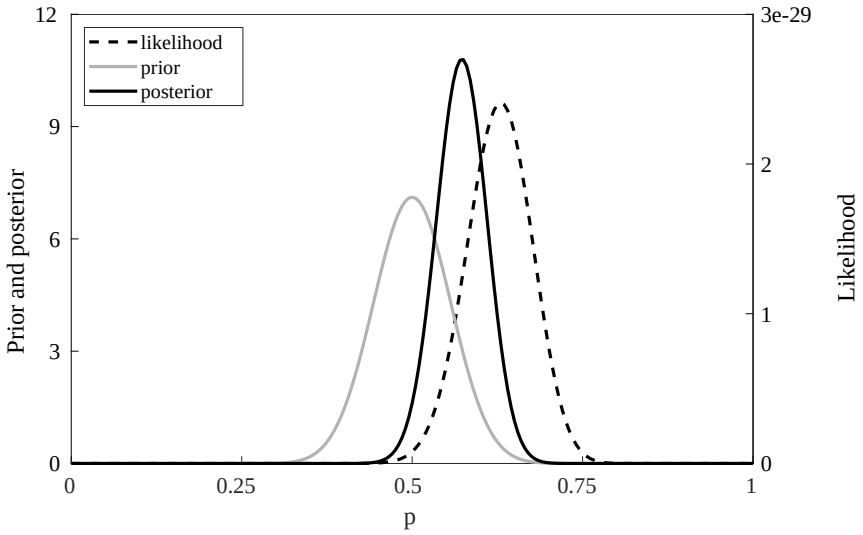


Figure 3.2: Likelihood, prior and posterior for the coin flip example

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:

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

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:

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

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.20 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.

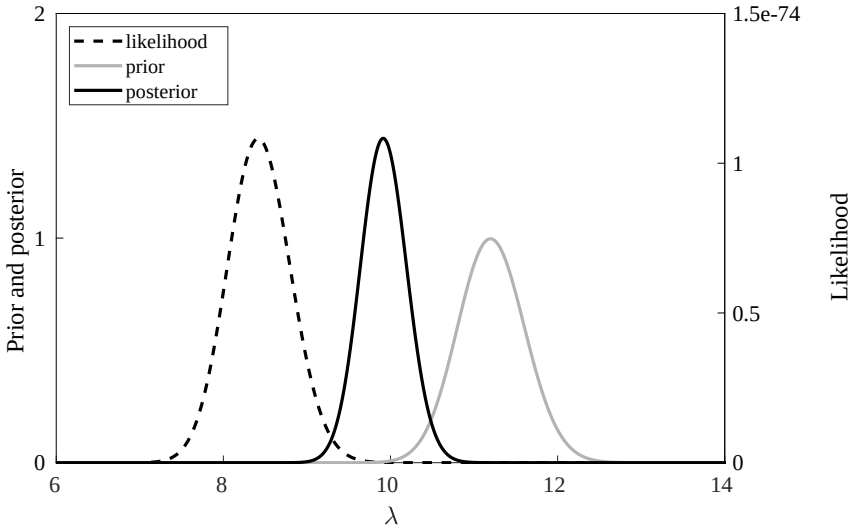


Figure 3.3: Likelihood, prior and posterior for the car sales example

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.

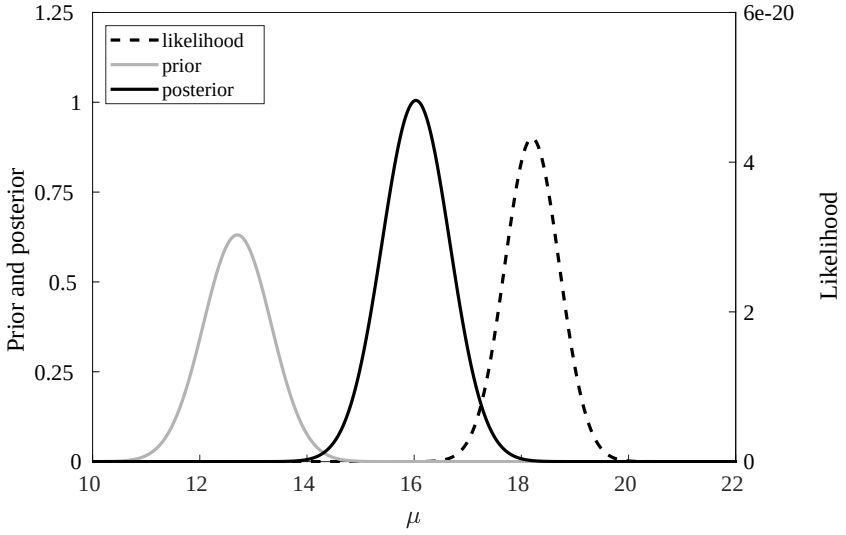


Figure 3.4: Likelihood, prior and posterior for the stock return example

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:

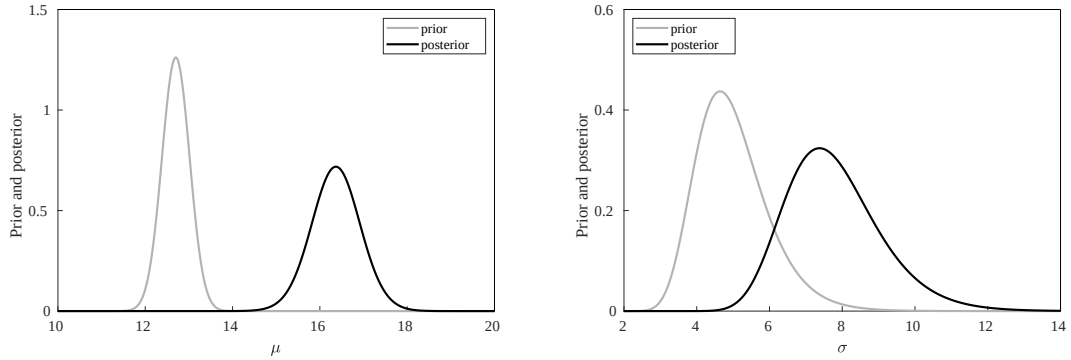
$$\pi(\mu|y) \propto \Gamma\left(\frac{\bar{\alpha}+1}{2}\right)\left(\frac{\bar{\delta}+(\mu-\bar{m})^2/\bar{v}}{2}\right)^{-\frac{\bar{\alpha}+1}{2}} \tag{1.4.9}$$

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.24 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.

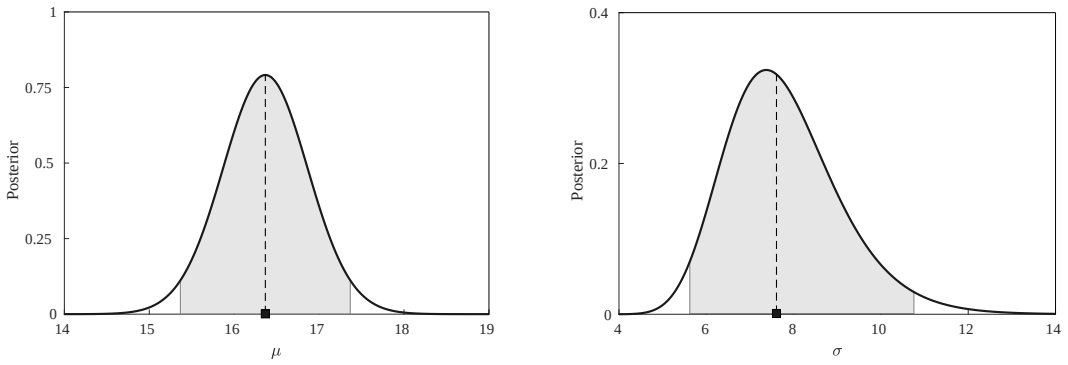
definition 4.5: let θ be some parameter; a **credibility interval** of level α is an interval of the form:
$\mathbb{P}(\theta_L \leq \theta \leq \theta_U|y) = 1-\alpha$
where θ_L and θ_U respectively denote the lower and upper bounds of the interval.

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:

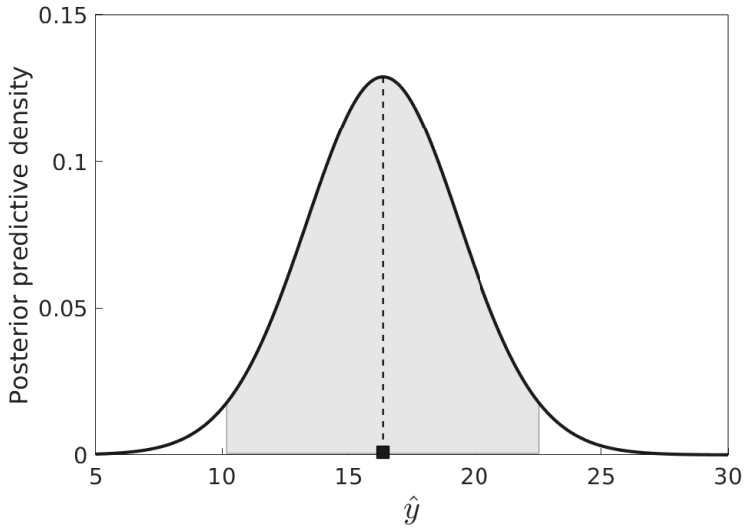


Figure 4.3: Posterior predictive distribution for the stock return example

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.

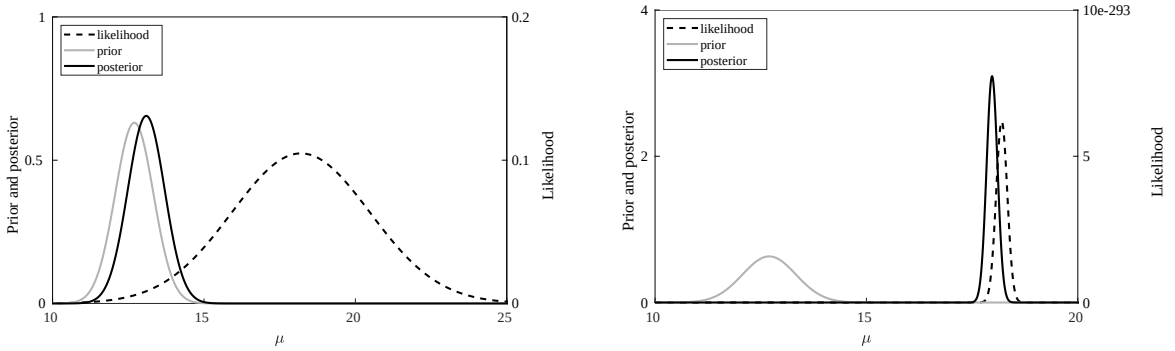
5.2 Large VS. small samples

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**.

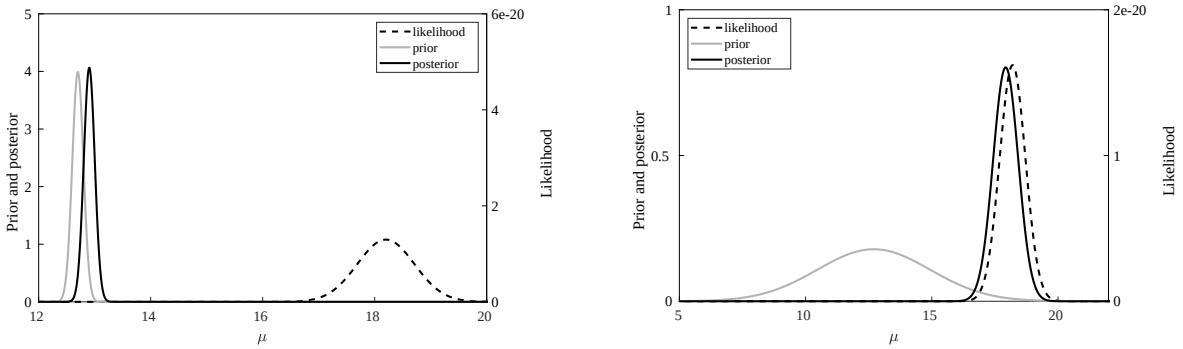
For instance, to specify a prior distribution on some parameter θ taking real values, we may use the improper prior $\pi(\theta) \propto 1$. This defines a uniform distribution over the interval $[-\infty, +\infty]$. The distribution integrates to infinity and connot be normalised to one. Improper priors 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 hierarachical 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:

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

3. at iteration j, draw:

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

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:

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

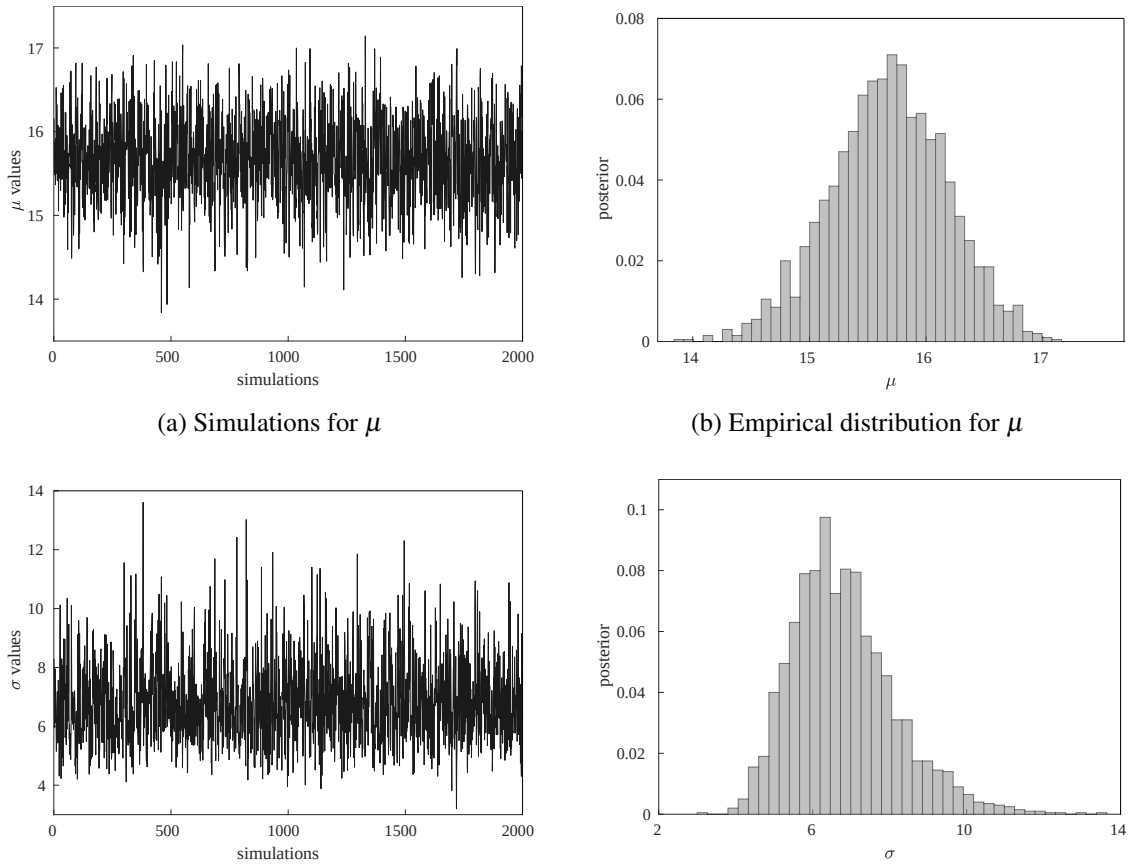
4. repeat to obtain 1000 iterations as burn-in sample and 2000 additional iterations for simulated values.

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 μ and σ

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:

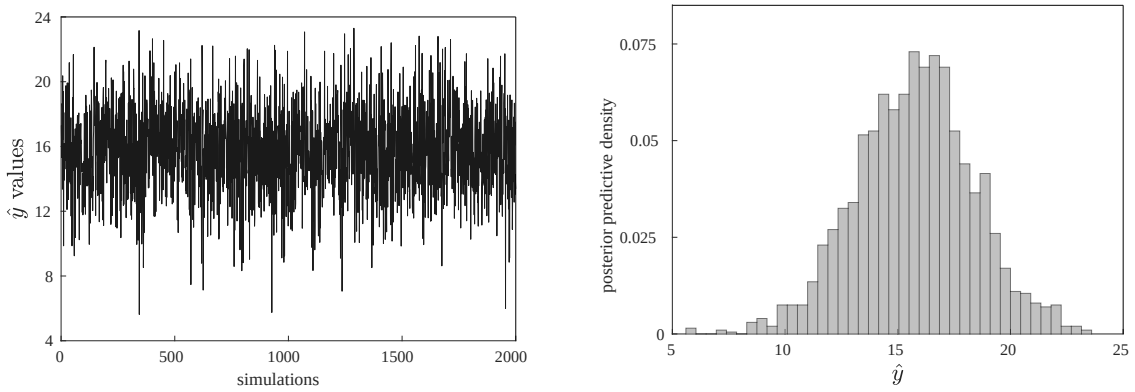
$$f(\hat{y}|y,\mu,\sigma) = (2\pi\sigma)^{-1/2}\exp\left(-\frac{1}{2}\frac{(\hat{y}-\mu)^2}{\sigma}\right) \tag{2.6.10}$$

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:

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

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:

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

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

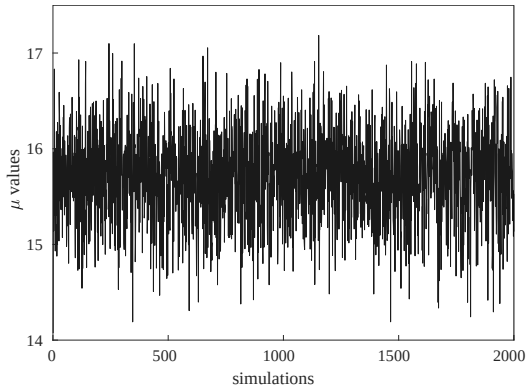
   if $u \leq \alpha(\theta_i^{(j-1)},\theta_i^{(j)})$, set $\theta_i^{(j)} = \tilde{\theta}_i$; else, set $\theta_i^{(j)} = \theta_i^{(j-1)}$

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

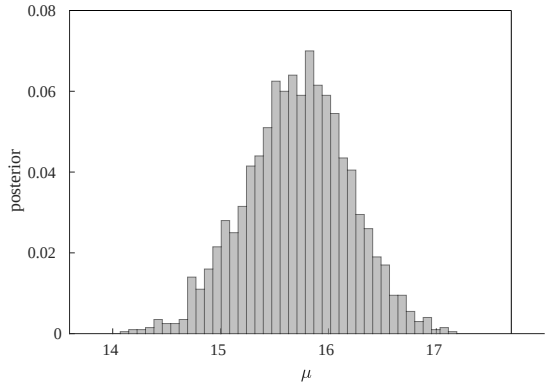
It remains to calibrate the prior $\pi(\lambda)$ and the transition kernel $q(\lambda^{(j-1)}, \lambda^{(j)})$. For $\pi(\lambda)$, we set $g = 1.6$ and $z = 0.04$. This way, the mean and variance of $\exp(\lambda)$ match the prior mean of 5 and the prior variance of 1 proposed for σ in section 4.3. 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.

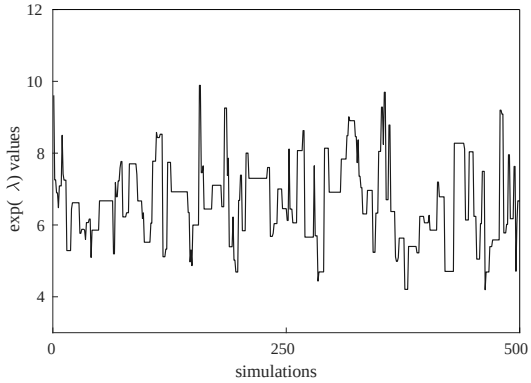
The bottom plots display the simulations and empirical distribution from the Metropolis-Hastings algorithm for λ. The left panel shows the first 500 iterations of the algorithm, before trimming is operated. The repeated values typical of the Metropolis-Hastings algorithm are quite apparent. The right 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.



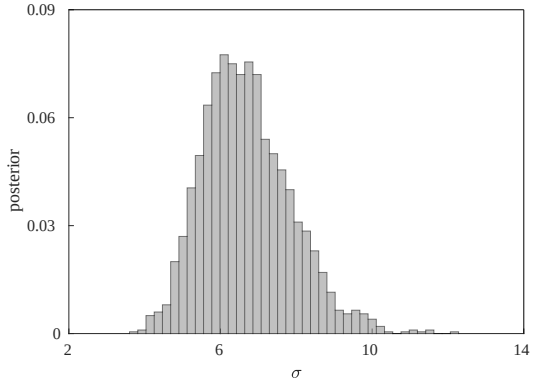
(a) Simulations for μ



(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):

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

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 (2008), 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.

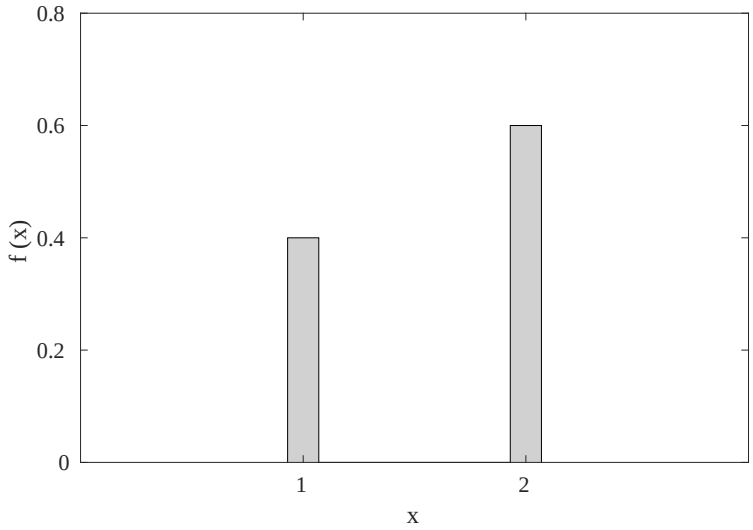


Figure 8.1: Probability mass function of the target distribution

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 $p_{ij} = \mathbb{P}(X_{t+1} = s_j | X_t = s_i)$, 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.

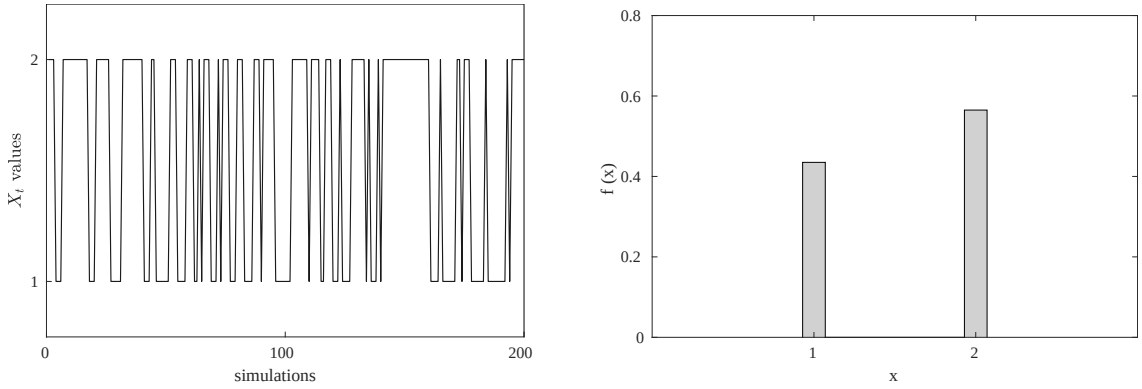
This suggests a natural procedure to generate values from a target finite distribution:

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:

$|p_{ij}^{(h)} - \pi_j| \leq \delta^{\,h/v}$ 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.

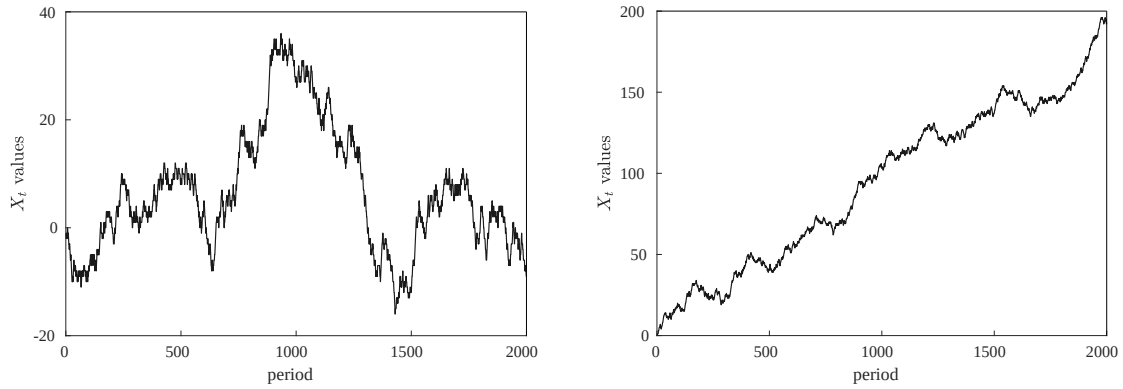
Understanding why the Markov chain has to be irreducible and aperiodic is straightforward. If the chain is not irreducible, then the exist at least two states s_i and s_j that don't communicate. In this case it is not possible to reach an invariant distribution since reaching state s_i precludes state 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 $\pi_i = \lim_{h\to+\infty} 1/h \sum_{t=1}^{h} \mathbb{1}(X_t = s_i)$

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 $\pi_j = \lim_{h\to+\infty} p_{ij}^{(h)}$ 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 $\pi_j = \lim_{h\to+\infty} p_{ij}^{(h)}$ 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:

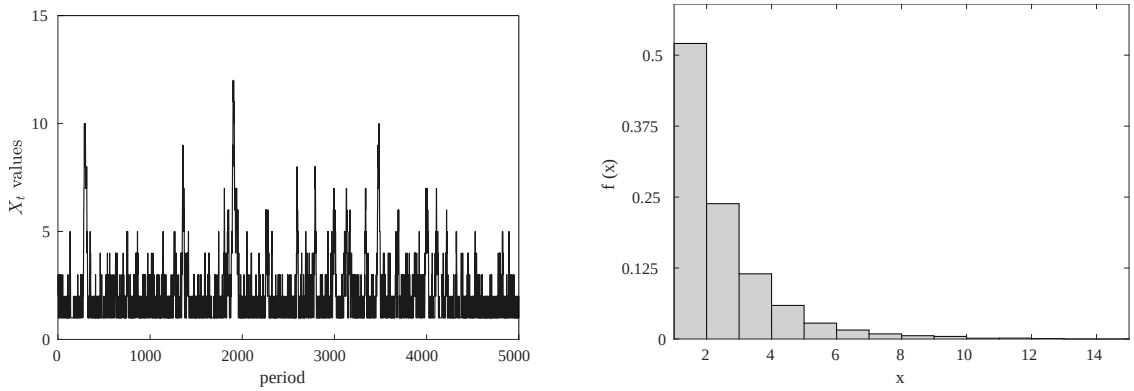
$$P = \begin{pmatrix} p+q & r & 0 & 0 & 0 & \cdots \\ p & q & r & 0 & 0 & \cdots \\ 0 & p & q & r & 0 & \cdots \\ 0 & 0 & p & q & r & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{2.8.10}$$

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 $\pi_j = \lim_{h\to+\infty} p_{ij}^{(h)}$: 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:

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

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 $A \in S$.

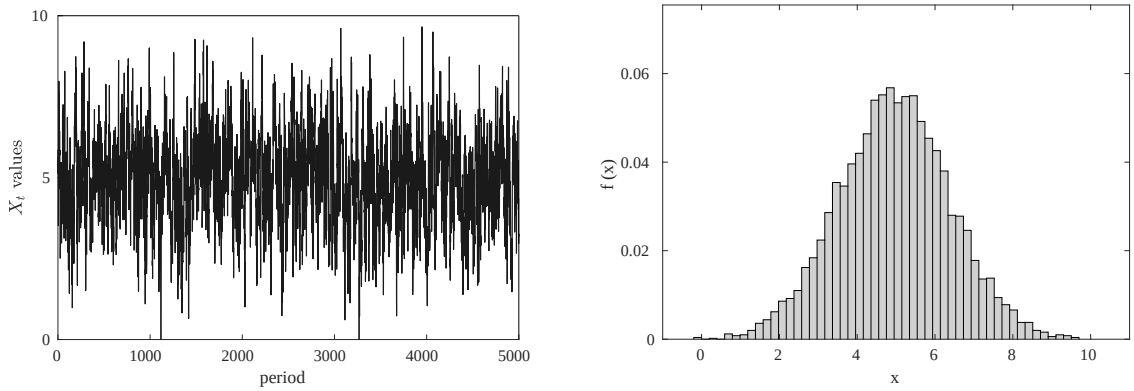
After some manipulations, this rewrites as (book 2, p. 26):

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

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:

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

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:

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

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:

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

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:

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

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

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

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:

$$\pi(\beta,\sigma|y) \propto f(y|\beta,\sigma)\pi(\beta)\pi(\sigma) \qquad (3.9.29)$$

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

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

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:

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

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:

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

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:

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

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:

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

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:

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

And equivalently, this can be done from:

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

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

$$\beta = G^{-1\prime}\left[G^{-1}(V^{-1}b + \sigma^{-1}X'y) + \zeta\right] \qquad \zeta \sim N(0,I_k) \tag{3.9.74}$$

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

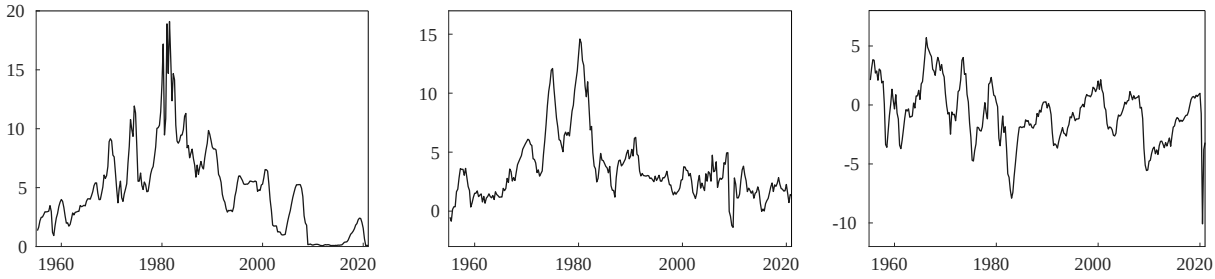
The conduct of monetary policy constitutes the core activity of central banking institutions. To understand how central institutions determine the leading interest rate, Taylor (1993) proposed a simple targetting rule linking the nominal interest rate to inflation and the output gap. Precisely, he postulated 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 $g = 0_h$, $Q = 100 I_h$, $p = 0_q$, and $Z = 100\, I_q$.

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:

$$AIC = 2k/n - 2\,\hat{L}/n \qquad BIC = k\log(n)/n - 2\,\hat{L}/n \tag{3.10.9}$$

where $\hat{L} = \log(f(y|\hat{\theta}))$ 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:

$$AIC = 2k/n + \log(\hat{\sigma}) \qquad BIC = k\,\log(n)/n + \log(\hat{\sigma}) \tag{3.10.10}$$

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 criteria 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 $\hat{s}_i = \sqrt{\sigma_i}$ 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 $f(\hat{y}_i|y) \sim T(\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i)$, and denote by $\tilde{y}_i = (y_i - \hat{\mu}_i)/\hat{s}_i$, where $\hat{s}_i = \sqrt{\sigma_i}$ is the square root of the scale parameter. Then the formula becomes:

$$CRPS = \hat{s}_i \left\{ \tilde{y}_i(2F(\tilde{y}_i) - 1) + 2f(\tilde{y}_i) \left(\frac{\hat{\nu}_i + \tilde{y}_i^2}{\hat{\nu}_i - 1} \right) - \frac{2\sqrt{\hat{\nu}_i}}{\hat{\nu}_i - 1} \frac{B(\frac{1}{2}, \hat{\nu}_i - \frac{1}{2})}{B(\frac{1}{2}, \frac{\hat{\nu}_i}{2})^2} \right\} \tag{3.10.16}$$

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:

$$CRPS \approx \frac{1}{J}\sum_{j=1}^{J} |\hat{y}_i^{(j)} - y_i| - \frac{1}{2J^2}\sum_{j=1}^{J}\sum_{k=1}^{J} |\hat{y}_i^{(j)} - \hat{y}_i^{(k)}| \tag{3.10.17}$$

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):

$$\log(f(y)) \approx -log\left((\omega J)^{-1}(2\pi)^{(n-1)/2}\,|\hat{\Sigma}|^{-1/2}\,|V|^{1/2}\,|Q|^{1/2}\;\frac{\Gamma(\alpha/2)}{\delta/2^{\alpha/2}}\right)$$
$$-log\left(\sum_{j=1}^{J}\mathbb{1}(\theta\in\hat{\Theta})\,|W|^{1/2}\,\sigma^{(\alpha+n)/2+1}\exp\left(\frac{1}{2}\left[\begin{array}{l}(y-X\beta)'(\sigma W)^{-1}(y-X\beta)+(\beta-b)'V^{-1}(\beta-b)\\+\delta\sigma^{-1}+(\gamma-g)'Q^{-1}(\gamma-g)-(\theta-\hat{\theta})'\hat{\Sigma}^{-1}(\theta-\hat{\theta})\end{array}\right]\right)\right) \tag{3.10.29}$$

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:

$$\frac{1}{f(y)} \approx \frac{1}{J}\sum_{j=1}^{J} \frac{g(\theta^{(j)})}{f(y|\beta^{(j)},\sigma^{(j)},\gamma^{(j)})\ \pi(\beta^{(j)})\ \pi(\sigma^{(j)})\ \pi(\phi^{(j)})} \tag{3.10.31}$$

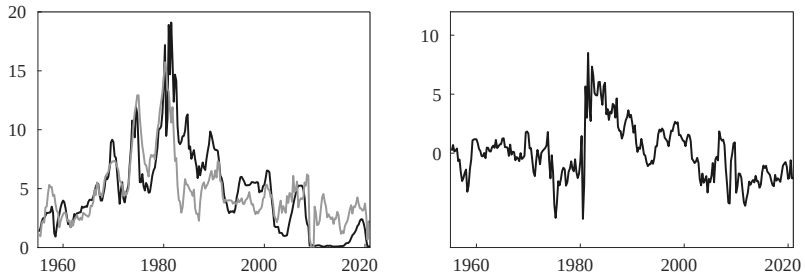
Using the probability density function (2.7.19) along with the likelihood function (3.9.61) and the priors (3.9.10), (3.9.21) and (3.9.62) then rearranging yields (book 2, p. 49):

$$\begin{aligned}\log(f(y)) \approx &-\log\left((\omega J)^{-1}(2\pi)^{(T-1)/2}\ |\hat{\Sigma}|^{-1/2}\ |V|^{1/2}\ |Z|^{1/2}\ \frac{\Gamma(\alpha/2)}{\delta/2^{\alpha/2}}\right)\\ -\log\Bigg(\sum_{j=1}^{J} \mathbb{1}(\theta\in\hat{\Theta})\sigma^{(\alpha+T)/2+1}\exp\Bigg(\frac{1}{2}\Bigg[\ &(\varepsilon-E\phi)'\sigma^{-1}(\varepsilon-E\phi)+(\beta-b)'V^{-1}(\beta-b)\\ &+\delta\sigma^{-1}+(\phi-p)'Z^{-1}(\phi-p)-(\theta-\hat{\theta})'\hat{\Sigma}^{-1}(\theta-\hat{\theta})\ \Bigg]\Bigg)\Bigg)\end{aligned} \tag{3.10.32}$$

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.

PART IV

Vector autoregressions

CHAPTER 11

Vector autoregressions

This chapter introduces the workhorse of Bayesian time-series econometrics: vector autoregressions. It focuses on model formulation and estimation. Additional aspects of the model will be discussed in subsequent chapters.

11.1 Formulation and maximum likelihood estimate

A general **vector autoregression** model, or VAR in short, can be formulated as:

$$y_t = Cz_t + A_1 y_{t-1} + \cdots + A_p y_{t-p} + \varepsilon_t \qquad \varepsilon_t \sim N(0,\Sigma) \qquad t = 1,\cdots,T \tag{4.11.1}$$

where y_t is a n-dimensional vector of **endogenous variables**, z_t is a m-dimensional vector of **exogenous variables** such as constant and trends, and ε_t is a n-dimensional vector of **residuals**. $A_1,\cdots,A_p$ are $n \times n$ coefficient matrices on the lagged values of y_t while C is a $n \times m$ matrix of coefficients on the exogenous regressors z_t. The n-dimensional vector of disturbance ε_t is assumed to be normally distributed with zero mean and variance-covariance matrix $\mathbb{E}(\varepsilon_t'\varepsilon_t) = \Sigma$, where Σ is symmetric and positive definite. The disturbances are non-autocorrelated so that so that $\mathbb{E}(\varepsilon_t'\varepsilon_s) = 0$ if $t \neq s$. The sample is observed over $t = 1,\cdots,T$ time periods.

A VAR model presents two main characteristics. It is **multivariate** as it represents a system of n simultaneous equations where each of the n endogenous variables is explained by itself and the other variables of the system. It is also **dynamic** since the variables y_t are explained not only by the exogenous regressors z_t, but also by their own lagged values $y_{t-1},\cdots,y_{t-p}$.

The VAR implies the estimation of $k = m + np$ coefficients for each equation, and thus a total of $q = nk$ coefficients for the full model. Estimation can be made more convenient by rewriting the VAR in compact form. Transposing and stacking the elements in (4.11.1) over the T sample periods yields:

$$Y = X\mathcal{B} + \mathcal{E} \tag{4.11.2}$$

with:

$$Y = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_T' \end{pmatrix} \qquad X = \begin{pmatrix} z_1' & y_0' & \cdots & y_{1-p}' \\ z_2' & y_1' & \cdots & y_{2-p}' \\ \vdots & \vdots & \ddots & \vdots \\ z_T' & y_{T-1}' & \cdots & y_{T-p}' \end{pmatrix} \qquad \mathcal{B} = \begin{pmatrix} C' \\ A_1' \\ \vdots \\ A_p' \end{pmatrix} \qquad \mathcal{E} = \begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_T' \end{pmatrix} \tag{4.11.3}$$

Y, X, $\mathcal{B}$ and $\mathcal{E}$ are matrices of respective dimensions $T \times n$, $T \times k$, $k \times n$ and $T \times n$. In practice, it is often easier to work with a vectorized version of (4.11.2). Using property m.54 one obtains:

$$y = \bar{X}\beta + \varepsilon \qquad \varepsilon \sim N(0,\bar{\Sigma}) \tag{4.11.4}$$

with:

$$y = vec(Y) \qquad \bar{X} = I_n \otimes X \qquad \beta = vec(\mathcal{B}) \qquad \varepsilon = vec(\mathcal{E}) \qquad \bar{\Sigma} = \Sigma \otimes I_T \tag{4.11.5}$$

y and ε are vectors of dimension nT while $\bar{X}$ is a $nT \times q$ matrix of regressors. β is a q-dimensional vector that gathers the dynamic coefficients of the model. With this formulation, one can define the parameters of interest for the model as $\theta = \{\beta, \Sigma\}$.

Consider first a frequentist approach of the VAR model. Following section 3.1, we need to set the likelihood function for the model. It follows immediately from (4.11.4) that $y \sim N(\bar{X}\beta, \bar{\Sigma})$. The likelihood function is then given by:

$$f(y|\beta,\Sigma) = (2\pi)^{-nT/2}|\bar{\Sigma}|^{-1/2} exp\left(-\frac{1}{2}(y-\bar{X}\beta)'\bar{\Sigma}^{-1}(y-\bar{X}\beta)\right) \tag{4.11.6}$$

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

$$\hat{\beta},\hat{\Sigma} = \underset{\beta,\Sigma}{argmax}\ \log(f(y|\beta,\Sigma)) \tag{4.11.7}$$

The log-likelihood function is given by:

$$\log(f(y|\beta,\Sigma)) = -\frac{nT}{2}\log(2\pi) - \frac{1}{2}\log(|\bar{\Sigma}|) - \frac{1}{2}(y-\bar{X}\beta)'\bar{\Sigma}^{-1}(y-\bar{X}\beta) \tag{4.11.8}$$

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

It can be shown (book 2, p. 53) that the resulting estimates are:

$$\hat{\beta} = vec(\hat{\mathcal{B}}) \qquad \hat{\mathcal{B}} = (X'X)^{-1}X'Y \qquad \hat{\Sigma} = \frac{1}{T}(Y - X\hat{\mathcal{B}})'(Y - X\hat{\mathcal{B}}) \tag{4.11.9}$$

The maximum likelihood estimates $\hat{\beta}$ and $\hat{\Sigma}$ for the VAR model can be seen to coincide with standard OLS estimates, save for a bias in $\hat{\Sigma}$ (the divisor should $T-k-1$ instead of T to obtain an unbiased estimator).

The maximum likelihood estimates are consistent. Thus for large samples a confidence interval at the α confidence level for any individual coefficient β_i can be obtained from (see for instance Hamilton (1994), chapter 11):

$$\hat{\beta}_i \pm N_{\alpha/2}\, s_i \qquad s_i = \sqrt{Q_{ii}} \qquad Q = \hat{\Sigma} \otimes (X'X)^{-1} \tag{4.11.10}$$

11.2 The Minnesota prior

This section introduces the simplest Bayesian VAR model, based on the prior initially developed by Litterman (1985) and Doan et al. (1984) at the University of Minnesota. In this version of the model, the residual variance-covariance matrix Σ is assumed to be known so that only the VAR coefficients β remain to be estimated. To do so, we define $\Sigma = \hat{\Sigma}$, the maximum likelihood estimate obtained from (4.11.9). In this case, we are left with $\theta = \{\beta\}$. From Bayes rule 3.3, the posterior $\pi(\beta|y)$ is given by:

$$\pi(\beta|y) \propto f(y|\beta)\pi(\beta) \tag{4.11.11}$$

The likelihood function $f(y|\beta)$ is given by (4.11.6). Consider then the prior distribution for β. The multivariate normal distribution appears as a natural 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:

$$\pi(\beta) = (2\pi)^{-q/2}|V|^{-1/2}\exp\left(-\frac{1}{2}(\beta-b)'V^{-1}(\beta-b)\right) \tag{4.11.12}$$

The definition of b and V represents a key aspect of Bayesian VAR modelling and will be discussed in details shortly. For now we treat these values as given and proceed with estimation: substituting the likelihood function (4.11.6) and the prior (4.11.12) in Bayes rule (4.11.11) and relegating to the normalization constant any term not involving β, one obtains:

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2}(y-\bar{X}\beta)'\bar{\Sigma}^{-1}(y-\bar{X}\beta)\right) \times \exp\left(-\frac{1}{2}(\beta-b)'V^{-1}(\beta-b)\right) \tag{4.11.13}$$

Starting from (4.11.13), rearranging and completing the squares, it can be shown (book 2, p. 54) that the posterior rewrites as:

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right) \tag{4.11.14}$$

with:

$$\bar{V} = (V^{-1}+\Sigma^{-1}\otimes X'X)^{-1} \qquad \bar{b} = \bar{V}(V^{-1}b + vec(X'Y\Sigma^{-1})) \tag{4.11.15}$$

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

Now that the posterior is derived, the question that remains is how b and V should be defined. This is a key element since the quality of the prior determines the relevance of the posterior. The strategy followed here is due to Litterman (1985) and has by now become canonical under the name of **Minnesota prior**.

The prior mean b postulates that most economic variables behave as random walks. Therefore, each variable included in a VAR model should be characterized by a value of 1 in its first own lag, and a value of 0 for any other coefficient. In practice, VAR models are often estimated with stationary variables. In this case, a value close to but smaller than 1 such as 0.95 may be prefered for the prior on the first own lag. As an example, consider a small VAR with 2 variables, 2 lags and one constant, and assume that each endogenous variable has its own autoregressive coefficient δ_i on its own first lag. Then one obtains:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \quad \Rightarrow \quad b = \begin{pmatrix} 0 \\ \delta_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \delta_2 \\ 0 \\ 0 \end{pmatrix} \tag{4.11.16}$$

For the prior covariance V, the strategy postulates a diagonal matrix (no prior covariance between coefficients) with smaller prior variance for coefficients on further lags and other variables, reflecting the belief that such coefficients are more likely to be equal to 0. On the contrary, a large prior variance is set on exogenous variables as little is known about their coefficient values. This gives the following cases:

1. Coefficients in β relating endogenous variables to their own lags. The prior variance is then given by:

$$\left(\frac{\pi_1}{lag^{\pi_3}}\right)^2 \tag{4.11.17}$$

where π_1 is an overall tightness parameter that applies to all coefficients, and π_3 is a parameter that controls the speed at which the prior variance on further lags is shrunk to 0.

2. Coefficients in β relating endogenous variables to the lags of other endogenous variables. The prior variance is then given by:

$$\left(\frac{s_i}{s_j}\right)\left(\frac{\pi_1\pi_2}{lag^{\pi_3}}\right)^2 \tag{4.11.18}$$

where s_i and s_j denote the residual variance of autoregressive models estimated by OLS for variables i and j, i being the explained variable and j the explanatory variable. π_2 is a cross-variable shrinkage parameter that further reduces the prior variance.

3. Coefficients in β related to exogenous variables. The prior variance is then given by:

$$s_i(\pi_1\pi_4)^2 \tag{4.11.19}$$

where π_4 is an exogenous-specific shrinkage parameter.

The following parameter values are commonly found in the litterature: $\pi_1 = 0.1$, $\pi_2 = 0.5$, $\pi_3 = 1$ or 2 and $\pi_4 = 100$ or more.

For the above simple VAR with 2 variables and 2 lags, the prior variance settings yields:

$$V = \begin{pmatrix}
s_1(\pi_1\pi_4)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & (\pi_1)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{s_1}{s_2}(\pi_1\pi_2)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \left(\frac{\pi_1}{2^{\pi_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{s_1}{s_2}(\frac{\pi_1\pi_2}{2^{\pi_3}})^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & s_2(\pi_1\pi_4)^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{s_2}{s_1}(\pi_1\pi_2)^2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & (\pi_1)^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{s_2}{s_1}(\frac{\pi_1\pi_2}{2^{\pi_3}})^2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\pi_1}{2^{\pi_3}}\right)^2
\end{pmatrix} \tag{4.11.20}$$

11.3 The Normal-Wishart prior

The main shortcoming of the original Minnesota prior is that it assumes that the residual variance-covariance matrix Σ is known. This is a strong assumption, and for this reason the Minnesota prior has been later updated to include Σ in the set of estimated parameters. The version developed in this section closely follows the presentation of Kadiyala and Karlsson (1997)[1].

The beauty of the Normal-Wishart prior resides in the existence of analytical posteriors despite the estimation of multiple parameters. This comes however at the cost of a hierarachical prior, the use of matricial distributions, and additional complexity in calculations.

In this setting, the model is thus still the VAR model introduced in (4.11.1), but the parameters of interest are extended to $\theta = \{\beta, \Sigma\}$. Following, Bayes rule is given by:

$$\pi(\beta, \Sigma|y) \propto f(y|\beta, \Sigma)\pi(\beta, \Sigma) \tag{4.11.21}$$

[1]The prior developed in this section is usually known as the "Normal-Wishart" prior. However, as it represents a late variation of the Minnesota prior, the litterature sometimes also designates it as the "Minnesota" prior. There is thus a possible ambiguity with the original Minnesota prior that must be settled from the context.

The Normal-Wishart builds a hierarchical prior by assuming that the prior distribution of β depends on the residual covariance matrix Σ. This construction is necessary to derive analytical posteriors. Following, we state $\pi(\beta,\Sigma) = \pi(\beta|\Sigma)\pi(\Sigma)$ and Bayes rule (4.11.21) rewrites:

$$\pi(\beta,\Sigma|y) \propto f(y|\beta,\Sigma)\ \pi(\beta|\Sigma)\ \pi(\Sigma) \tag{4.11.22}$$

The likelihood function is still given by (4.11.6). For later calculations, it is however convenient to reformulate it: after some manipulations, it can be shown (book 2, p. 55) that it rewrites as:

$$\begin{aligned} f(y|\beta,\Sigma) = \ & (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left(-\frac{1}{2}(\beta-\hat{\beta})'\ (\Sigma\otimes(X'X)^{-1})^{-1}\ (\beta-\hat{\beta})\right) \\ & \times exp\left(-\frac{1}{2}\ tr\left[\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right]\right) \end{aligned} \tag{4.11.23}$$

with $Y, X, \beta, \hat{B}$ and $\hat{\beta}$ defined as in (4.11.2) and (4.11.9). The first row of (4.11.23) can be recognised as the kernel of a multivariate normal distribution for β, and the second row as the kernel of an inverse Wishart distribution for Σ, both centered around maximum likelihood estimates. It then seems natural to assume the same prior distributions for β and Σ to obtain conjugacy, and indeed this is the strategy that is going to be applied.

For β, the multivariate normal distribution represents again a good candidate. However, we note that the first row of (4.11.23) suggests a dependence of β on Σ through the Kronecker structure, which motivates the use of a hierarchical prior. The prior is thus defined as a multivariate normal distribution with prior mean b and prior variance $\Sigma\otimes W$: $\pi(\beta|\Sigma) \sim N(b, \Sigma\otimes W)$:

$$\pi(\beta|\Sigma) = (2\pi)^{-q/2}|\Sigma\otimes W|^{-1/2}\exp\left(-\frac{1}{2}(\beta-b)'(\Sigma\otimes W)^{-1}(\beta-b)\right) \tag{4.11.24}$$

The prior mean b is defined similarly to the Minnesota prior. For the prior variance, note the difference between W in (4.11.24) and V in the Minnesota prior (4.11.12): while V represents the full variance-covariance matrix of β, W only represents the prior variance of a single equation in the VAR. The specific Kronecker structure of the hierarchical prior then implies that each equation has its prior variance scaled by the residual variance given by Σ. To keep the overall structure as close as possible to the original Minnesota prior, W is defined as follows:

1. For the coefficients in β relating endogenous variables to their own lags or to the lags of other endogenous variables, the prior variance is given by:

$$\left(\frac{1}{s_j}\right)\left(\frac{\pi_1}{lag^{\pi_3}}\right)^2 \tag{4.11.25}$$

where π_1 and π_3 are the overall tightness and lag decay parameters of the Minnesota prior, and s_j denotes again the residual variance of an autoregressive models estimated by OLS for explanatory variable j.

2. For the coefficients in β related to exogenous variables, the prior variance is given by:

$$(\pi_1\pi_4)^2 \tag{4.11.26}$$

Consider again a simple VAR model with 2 variables and 2 lags. The prior matrix W then writes as:

$$W = \begin{pmatrix} (\pi_1\pi_4)^2 & 0 & 0 & 0 & 0 \\ 0 & \left(\frac{1}{s_1}\right)\pi_1^2 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{1}{s_2}\right)\pi_1^2 & 0 & 0 \\ 0 & 0 & 0 & \left(\frac{1}{s_1}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & \left(\frac{1}{s_2}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2 \end{pmatrix} \tag{4.11.27}$$

Following, the full prior covariance $\Sigma \otimes W$ matrix writes:

$$\Sigma \otimes W = \begin{pmatrix}
\sigma_{11}(\pi_1\pi_4)^2 & 0 & 0 & 0 & 0 & \sigma_{12}(\pi_1\pi_4)^2 & 0 & 0 & 0 & 0 \\
0 & \left(\frac{\sigma_{11}}{s_1}\right)\pi_1^2 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_{12}}{s_1}\right)\pi_1^2 & 0 & 0 & 0 \\
0 & 0 & \left(\frac{\sigma_{11}}{s_2}\right)\pi_1^2 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_{12}}{s_2}\right)\pi_1^2 & 0 & 0 \\
0 & 0 & 0 & \left(\frac{\sigma_{11}}{s_1}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_{12}}{s_1}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2 & 0 \\
0 & 0 & 0 & 0 & \left(\frac{\sigma_{11}}{s_2}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_{12}}{s_2}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2 \\
\sigma_{21}(\pi_1\pi_4)^2 & 0 & 0 & 0 & 0 & \sigma_{22}(\pi_1\pi_4)^2 & 0 & 0 & 0 & 0 \\
0 & \left(\frac{\sigma_{21}}{s_1}\right)\pi_1^2 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_{22}}{s_1}\right)\pi_1^2 & 0 & 0 & 0 \\
0 & 0 & \left(\frac{\sigma_{21}}{s_2}\right)\pi_1^2 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_{22}}{s_2}\right)\pi_1^2 & 0 & 0 \\
0 & 0 & 0 & \left(\frac{\sigma_{21}}{s_1}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_{22}}{s_1}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2 & 0 \\
0 & 0 & 0 & 0 & \left(\frac{\sigma_{21}}{s_2}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_{22}}{s_2}\right)\left(\frac{\pi_1}{2^{\pi_3}}\right)^2
\end{pmatrix} \tag{4.11.28}$$

Note the similarities and differences between the Minnesota covariance matrix (4.11.20) and the normal-Wishart matrix (4.11.28). The two matrices follow the same shrinkage pattern, applying overall and lag-specific shrinkage through the parameters π_1 and π_3. The specific Kronecker structure of the normal-Wishart prior precludes however the application of the cross-variable shrinkage π_2. This makes the normal-Wishart prior comparable to a Minnesota prior with $\pi_2 = 1$, a restrictive and intrinsically undesirable assumption. Also, we can see that the normal Whishart uses both σ_{ij} (entry i, j of Σ) and s_i for scaling. This makes the prior analysis more complex as the σ_{ij} terms are not constant hyperparameters but random variables endogenously estimated within the model. Finally, the off-diagonal terms imply that the normal-Wishart prior generates prior covariance between the coefficients, another potentially undesirable assumption.

It remains to define a prior distribution for the residual variance-covariance matrix Σ. The usual choice is an inverse Wishart distribution with degrees of freedom α and scale S: $\pi(\Sigma) \sim IW(\alpha, S)$. Following:

$$\pi(\Sigma) = \frac{2^{-\alpha n/2}}{\Gamma_n\left(\frac{\alpha}{2}\right)}|S|^{\alpha/2}|\Sigma|^{-(\alpha+n+1)/2}exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}S\right\}\right) \tag{4.11.29}$$

Kadiyala and Karlsson (1997) suggests to define α and S as:

$$\alpha = n+2 \qquad\qquad S = (\alpha - n - 1)\begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_n \end{pmatrix} \tag{4.11.30}$$

where s_i denotes again the residual variance of an autoregressive models estimated by OLS for explanatory variable i. The prior degrees of freedom α is set at the smallest value ensuring the existence of a prior mean and variance for Σ, keeping the distribution well-defined but as uninformative as possible. The prior scale is defined so that $\mathbb{E}(\Sigma) = diag(s_1, s_2, \cdots, s_n)$, in other words to replicate the elements of a diagonal OLS estimate.

From Bayes rule (4.11.22), we combine the likelihood function (4.11.23) with the priors (4.11.24) and (4.11.29) to obtain:

$$\begin{aligned}\pi(\beta,\Sigma|y) &\propto |\Sigma|^{-T/2} exp\left(-\frac{1}{2}(\beta-\hat{\beta})'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}(\beta-\hat{\beta})\right)\\ &\times exp\left(-\frac{1}{2}\,tr\left[\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right]\right)\\ &\times |\Sigma\otimes W|^{-1/2}\exp\left(-\frac{1}{2}(\beta-b)'(\Sigma\otimes W)^{-1}(\beta-b)\right)\\ &\times |\Sigma|^{-(\alpha+n+1)/2} exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}S\right\}\right)\end{aligned} \tag{4.11.31}$$

where we have relegated to the normalization constant any multiplicative term not involving β or Σ. This joint posterior can rewrite as the product of a matrix normal distribution for $\mathcal{B}$ (the unvectorized form of β in (4.11.2)) and an inverse Wishart for Σ. Indeed, it can be shown (book 2, p. 56) that:

$$\begin{aligned}\pi(\mathcal{B},\Sigma|y) &\propto |\Sigma|^{-k/2} exp\left(-\frac{1}{2}\,tr\left\{\Sigma^{-1}(\mathcal{B}'-\bar{B})'\bar{W}^{-1}(\mathcal{B}'-\bar{B})\right\}\right)\\ &\times |\Sigma|^{-(\bar{\alpha}+n+1)/2} exp\left(-\frac{1}{2}\,tr\left\{\Sigma^{-1}\bar{S}\right\}\right)\end{aligned} \tag{4.11.32}$$

with:

$$\bar{W}=(W^{-1}+X'X)^{-1} \qquad \bar{B}=\bar{W}(W^{-1}B+X'Y) \qquad \bar{\alpha}=\alpha+T \qquad \bar{S}=S+Y'Y+B'W^{-1}B-\bar{B}'\bar{W}^{-1}\bar{B} \tag{4.11.33}$$

and were B denotes the prior mean vector b reorganised as a $k\times n$ matrix.

This formulation eventually makes it possible to derive the marginal posteriors by using definition 4.3. Marginalization is easy for Σ as $\mathcal{B}$ only appears in the first part of the posterior:

$$\begin{aligned}\pi(\Sigma|y) &= \int \pi(\mathcal{B},\Sigma|y)d\mathcal{B}\\ &\propto \int |\Sigma|^{-k/2} exp\left(-\frac{1}{2}\,tr\left\{\Sigma^{-1}(\mathcal{B}'-\bar{B})'\bar{W}^{-1}(\mathcal{B}'-\bar{B})\right\}\right)d\mathcal{B}\\ &\times |\Sigma|^{-(\bar{\alpha}+n+1)/2} exp\left(-\frac{1}{2}\,tr\left\{\Sigma^{-1}\bar{S}\right\}\right)\\ &\propto |\Sigma|^{-(\bar{\alpha}+n+1)/2} exp\left(-\frac{1}{2}\,tr\left\{\Sigma^{-1}\bar{S}\right\}\right)\end{aligned} \tag{4.11.34}$$

where we have used the fact that the kernel of a density function integrates to a constant. This is recognized as the kernel of an inverse Wishart distribution with degrees of freedom $\bar{\alpha}$ and scale $\bar{S}$: $\pi(\Sigma|y)\sim IW(\bar{\alpha},\bar{S})$.

Obtaining the marginal for $\mathcal{B}$ is trickier. As Σ appears in both terms in (4.11.32), we group them and integrate:

$$\pi(\mathcal{B}|y)=\int \pi(\mathcal{B},\Sigma|y)d\Sigma \;\propto\; \int |\Sigma|^{-(\bar{\alpha}+k+n+1)/2} exp\left(-\frac{1}{2}\,tr\left\{\Sigma^{-1}\left[\bar{S}+(\mathcal{B}'-\bar{B})'\bar{W}^{-1}(\mathcal{B}'-\bar{B})\right]\right\}\right)d\Sigma \tag{4.11.35}$$

This is the kernel of an inverse Wishart distribution with degrees of freedom $(\bar{\alpha}+k)$ and scale $\bar{S}+(\mathcal{B}'-\bar{B})'\bar{W}^{-1}(\mathcal{B}'-\bar{B})$, and integration yields the reciprocal of the normalization constant of the distribution. Hence:

$$\pi(\mathcal{B}|y) \;\propto\; \Gamma_n\left(\frac{\bar{\alpha}+k}{2}\right)2^{(\bar{\alpha}+k)n/2}\left|\bar{S}+(\mathcal{B}'-\bar{B})'\bar{W}^{-1}(\mathcal{B}'-\bar{B})\right|^{-\frac{\bar{\alpha}+k}{2}} \;\propto\; \left|\bar{S}+(\mathcal{B}'-\bar{B})'\bar{W}^{-1}(\mathcal{B}'-\bar{B})\right|^{-\frac{\bar{\alpha}+k}{2}} \tag{4.11.36}$$

Finally, after some manipulations, it can be shown (book 2, p. 59) that this reformulates as:

$$\pi(\mathcal{B}|y) \propto \left| I_n + \frac{1}{\hat{\alpha}} \hat{S}^{-1} (\mathcal{B}' - \bar{B})' \bar{W}^{-1} (\mathcal{B}' - \bar{B}) \right|^{-\frac{\hat{\alpha}+k+n-1}{2}} \tag{4.11.37}$$

with:

$$\hat{\alpha} = \alpha + T - n + 1 \qquad \hat{S} = \bar{S}/\hat{\alpha} \tag{4.11.38}$$

This is the kernel of a matrix Student distribution with location $\bar{B}$, scales $\bar{W}$ and $\hat{S}$, and degrees of freedom $\hat{\alpha}$: $\pi(\mathcal{B}|y) \sim MT(\bar{B}, \bar{W}, \hat{S}, \hat{\alpha})$.

11.4 The independent prior

The normal-Wishart prior solves some of the shortcomings of the original Minnesota prior by estimating both β and Σ. This comes however at the cost of assuming prior dependence of β on Σ, an undesirable assumption, and fix hyperparameter value $\pi_2 = 1$. This section solves this issue by proposing a prior for β that is independent of Σ. The prior however does not admit analytical solutions and simulation methods are required.

In this setup, the model is again the VAR model introduced in (4.11.1), and the parameters of interest are $\theta = \{\beta, \Sigma\}$. Following, Bayes rule is still given by (4.11.21). However, assuming independence yields $\pi(\beta, \Sigma) = \pi(\beta)\ \pi(\Sigma)$ so that Bayes rule rewrites:

$$\pi(\beta, \Sigma|y) \propto f(y|\beta, \Sigma)\pi(\beta)\pi(\Sigma) \tag{4.11.39}$$

The likelihood function is unchanged and given by (4.11.6). The prior for β is the canonical Minnesota prior given by (4.11.12). For Σ, we use again the inverse Wishart prior given by (4.11.29). Substituting in Bayes rule (4.11.39) yields:

$$\begin{aligned} \pi(\beta, \Sigma|y) \ \ & |\bar{\Sigma}|^{-1/2} exp\left(-\frac{1}{2}(y - \bar{X}\beta)'\bar{\Sigma}^{-1}(y - \bar{X}\beta) \right) \\ & \times exp\left(-\frac{1}{2}(\beta - b)'V^{-1}(\beta - b) \right) \\ & \times |\Sigma|^{-(\alpha+n+1)/2} exp\left(-\frac{1}{2} tr\left\{ \Sigma^{-1} S \right\} \right) \end{aligned} \tag{4.11.40}$$

where as usual any multiplicative term not involving β or Σ has been relegated to the normalization constant. Unlike the normal-Wishart prior, there is no way here to integrate out the joint posterior (4.11.40) to obtain the marginal posteriors $\pi(\beta|y)$ and $\pi(\Sigma|y)$. The parameters β and Σ are too interwoven to permit integration. The only possibility then consists in using simulation methods and rely on 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 (4.11.40) and relegating to the normalization constant any multiplicative term not involving β, yielding:

$$\pi(\beta|y, \Sigma) \propto exp\left(-\frac{1}{2}(y - \bar{X}\beta)'\bar{\Sigma}^{-1}(y - \bar{X}\beta) \right) \times \exp\left(-\frac{1}{2}(\beta - b)'V^{-1}(\beta - b) \right) \tag{4.11.41}$$

This expression is similar to (4.11.13), and thus a similar procedure of rearranging and completing the squares yields:

$$\pi(\beta|y,\Sigma) \propto \exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right) \tag{4.11.42}$$

with:

$$\bar{V} = (V^{-1}+\Sigma^{-1}\otimes X'X)^{-1} \qquad \bar{b} = \bar{V}(V^{-1}b + vec(X'Y\Sigma^{-1})) \tag{4.11.43}$$

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

Obtain next the conditional posterior $\pi(\Sigma|y,\beta)$. Start from the joint posterior (4.11.40) and relegate to the normalization constant any multiplicative term not involving Σ to obtain:

$$\pi(\Sigma|y,\beta) \propto |\bar{\Sigma}|^{-1/2} exp\left(-\frac{1}{2}(y-\bar{X}\beta)'\bar{\Sigma}^{-1}(y-\bar{X}\beta)\right) \times |\Sigma|^{-(\alpha+n+1)/2} exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}S\right\}\right) \tag{4.11.44}$$

After rearranging, it can be shown (book 2, p. 59) that this expression rewrites:

$$\pi(\Sigma|y,\beta) \propto |\Sigma|^{-(\bar{\alpha}+n+1)/2}\ exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}\bar{S}\right\}\right) \tag{4.11.45}$$

with:

$$\bar{\alpha} = \alpha + T \qquad \bar{S} = S + (Y - X\mathcal{B})'(Y - X\mathcal{B}) \tag{4.11.46}$$

This is the kernel of an inverse Wishart distribution with degrees of freedom $\bar{\alpha}$ and scale $\bar{S}$: $\pi(\Sigma|y,\beta) \sim IW(\bar{\alpha},\bar{S})$.

We can then introduce the Gibbs sampling algorithm for the independent prior.

algorithm 11.1: Gibbs sampling algorithm for the VAR model with independent prior

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:

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

   $\Sigma^{(j)}$ from $\pi(\Sigma|y,\beta) \sim IW(\bar{\alpha},\bar{S})$ with:

   $\bar{\alpha} = \alpha + T \qquad \bar{S} = S + (Y - X\mathcal{B})'(Y - X\mathcal{B})$
4. repeat until the desired number of iterations is realised.

11.5 The dummy observation prior

The Minnesota and independent priors represent standard approaches to Bayesian VAR modelling. These two priors may however become unworkable for large models. Indeed, they involve the inversion of a $q \times q$ matrix $\bar{V}$, where $q = n(m+np)$ renders the inversion cost prohibitive when the number of endogenous variables n gets large.

One possible solution consists in using the normal-Wishart prior which only involves inverting a $k \times k$ matrix $\bar{W}$ with $k = m+np$, and thus reduces the dimensionality of the inversion by a factor n. Banbura et al. (2010) propose an equivalent solution that uses **dummy observations** to replicate the normal-Wishart

prior while improving further the computational efficiency of the procedure. In this section we first introduce the efficient prior, and then show how this prior can replicate the normal-Wishart prior by the way of dummy observations.

So, consider the VAR model introduced in (4.11.1) with parameters of interest $\theta = \{\beta, \Sigma\}$. Bayes rule is still given by (4.11.21), and assuming independence yields again $\pi(\beta, \Sigma) = \pi(\beta)\ \pi(\Sigma)$. Thus that Bayes rule rewrites:

$$\pi(\beta, \Sigma|y) \propto f(y|\beta, \Sigma)\pi(\beta)\pi(\Sigma) \tag{4.11.47}$$

For the likelihood function, it is convenient to use the normal-Wishart reformulation (4.11.23). This likelihood function is then combined with uninformative prior for β and Σ. Specifically, we use the so-called Jeffrey's priors proposed by Zellner (1996):

$$\pi(\beta) \propto 1 \qquad \pi(\Sigma) \propto |\Sigma|^{-(\alpha+1)/2} \tag{4.11.48}$$

with α defined in (4.11.30).These are the least possible informative priors[2]. There are two benefits from using such priors. First, they don't carry prior information, which is desirable since the prior beliefs will be conveyed through the dummy observations, as detailed below. Second, these priors make the form of the posterior extremely simple, which maximizes the computational efficiency of the model. Combining the likelihood function (4.11.23) with the priors (4.11.48) and rearranging, the joint posterior can be shown (book 2, p. 60) to write as:

$$\begin{aligned} \pi(\beta, \Sigma|y) \ &\propto |\Sigma|^{-k/2}\ exp\left(-\frac{1}{2}\ tr\left\{\Sigma^{-1}(\mathcal{B}' - \hat{B})'\hat{W}^{-1}(\mathcal{B}' - \hat{B})\right\}\right) \\ &\times |\Sigma|^{-(\hat{\alpha}+n+1)/2}\ exp\left(-\frac{1}{2}\ tr\left\{\Sigma^{-1}\hat{S}\right\}\right) \end{aligned} \tag{4.11.49}$$

with:

$$\hat{W} = (X'X)^{-1} \qquad \hat{\alpha} = T - k + 2 \qquad \hat{S} = (Y - X\hat{B})'(Y - X\hat{B}) \qquad \hat{B} = (X'X)^{-1}XY \tag{4.11.50}$$

Expectedly, using uninformative priors leaves the data as the only source of information and thus results in a posterior distribution centered around maximum likelihood estimates. (4.11.49) can be seen as the product of a matrix normal distribution for $\mathcal{B}$ centered around the OLS estimates $\hat{B}$, and an inverse Wishart distribution for Σ centered on the OLS sum of squared residuals $\hat{S}$.

The marginal posteriors are derived by repeating the same procedure as for the normal-Wishart prior. The marginal posterior $\pi(\Sigma|y)$ obtains from (4.11.49), following the same steps as (4.11.34). This directly yields:

$$\pi(\Sigma|y) \propto |\Sigma|^{-(\hat{\alpha}+n+1)/2}\ exp\left(-\frac{1}{2}\ tr\left\{\Sigma^{-1}\hat{S}\right\}\right) \tag{4.11.51}$$

This the kernel of an inverse Wishart distribution with degrees of freedom $\hat{\alpha}$ and scale $\hat{S}$: $\pi(\Sigma|y) \sim IW(\hat{\alpha}, \hat{S})$.

The marginal posterior $\pi(\mathcal{B}|y)$ obtains by following the same steps as (4.11.35) - (4.11.38). It can then be shown (book 2, p. 60) that:

$$\pi(\mathcal{B}|y) \propto \left| I_n + \frac{1}{\tilde{\alpha}}\tilde{S}^{-1}(\mathcal{B}' - \hat{B})'\hat{W}^{-1}(\mathcal{B}' - \hat{B}) \right|^{-\frac{\tilde{\alpha}+k+n-1}{2}} \tag{4.11.52}$$

[2]The traditional definition of Jeffrey's prior for Σ is $\pi(\Sigma) \propto |\Sigma|^{-(n+1)/2}$. Using instead $\pi(\Sigma) \propto |\Sigma|^{-(\alpha+1)/2}$ here maintains exact consistency with the Normal-Wishart prior.

with:

$$\tilde{\alpha} = T - n - k + 3 \qquad \tilde{S} = \hat{S}/\tilde{\alpha} \tag{4.11.53}$$

This is the kernel of a matrix Student distribution with location $\hat{B}$, scales $\hat{W}$ and $\tilde{S}$, and degrees of freedom $\tilde{\alpha}$: $\pi(\mathcal{B}|y) \sim MT(\hat{B}, \hat{W}, \tilde{S}, \tilde{\alpha})$.

The efficiency of the prior becomes apparent from (4.11.49) and (4.11.52). The posterior $\pi(\mathcal{B}|y)$ only requires the inversion of the $k \times k$ matrix $\hat{W}$, similarly to the normal-Wishart prior. Additionaly, all the posterior parameters only involve maximum likelihood estimates, which are the fastest to compute.

The downside of this prior is that it results in a posterior that is hardly different from a maximum likelihood estimate. If one provides no prior information whatsoever, there is, in fact, little point in using a Bayesian approach. To remedy this situation, one would ideally include prior beliefs in the model despite the uninformative priors. This can be done by the way of dummy, or artifical observations. These artifical observations are added to the actual sample observations in order to match the normal-Wishart prior distribution.

Precisely, we define the following matrices of artificial data:

$$Y_{dum} = \begin{pmatrix} 0_{m\times n} \\ \hline H/\pi_1 \\ 0_{n(p-1)\times n} \\ \hline K \end{pmatrix} \qquad X_{dum} = \begin{pmatrix} I_m/(\pi_1\pi_4) & 0_{m\times np} \\ \hline 0_{np\times m} & J\otimes K/\pi_1 \\ \hline 0_{n\times m} & 0_{n\times np} \end{pmatrix}$$
$$H = diag(\delta_1\sqrt{s_1}, \cdots, \delta_n\sqrt{s_n}) \qquad J = diag(1^{\pi_3}, 2^{\pi_3}, \cdots, p^{\pi_3}) \qquad K = diag(\sqrt{s_1}, \cdots, \sqrt{s_n}) \tag{4.11.54}$$

where $\delta_i, s_i, \pi_1, \pi_3$ and π_4 are defined similarly to the Minnesota prior. Y_d and X_d are of dimension $(n(p+1)+m) \times n$ and $(n(p+1)+m) \times (m+np)$, respectively. Each row of these matrices corresponds to one dummy observation so that we create $T_{dum} = n(p+1)+m$ artifical observations. The dummy observations are divided in three blocks. The first and second block implement the Minnesota prior for the exogenous and endogenous variables respectively, while the third block corresponds to the prior belief for the residual covariance matrix.

To understand how these observations replicate the normal-Wishart prior, consider again a simple VAR model with 2 variables, 2 lags and a constant. Then use the stacked form (4.11.2) of the VAR, applied to the artificial observations.

$$Y_{dum} = X_{dum}\mathcal{B} + \mathcal{E}_{dum} \tag{4.11.55}$$

Using the matrix definitions in (4.11.48), this yields:

$$\begin{pmatrix} 0 & 0 \\ \hline \frac{\delta_1\sqrt{s_1}}{\pi_1} & 0 \\ 0 & \frac{\delta_2\sqrt{s_2}}{\pi_1} \\ 0 & 0 \\ 0 & 0 \\ \hline \sqrt{s_1} & 0 \\ 0 & \sqrt{s_2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\pi_1\pi_4} & 0 & 0 & 0 & 0 \\ \hline 0 & \frac{\sqrt{s_1}1^{\pi_3}}{\pi_1} & 0 & 0 & 0 \\ 0 & 0 & \frac{\sqrt{s_2}1^{\pi_3}}{\pi_1} & 0 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{s_1}2^{\pi_3}}{\pi_1} & 0 \\ 0 & 0 & 0 & 0 & \frac{\sqrt{s_2}2^{\pi_3}}{\pi_1} \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} c_{11} & c_{21} \\ a_{11}^1 & a_{21}^1 \\ a_{12}^1 & a_{22}^1 \\ a_{11}^2 & a_{21}^2 \\ a_{12}^2 & a_{22}^2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{2,1} \\ \hline \varepsilon_{1,2} & \varepsilon_{2,2} \\ \varepsilon_{1,3} & \varepsilon_{2,3} \\ \varepsilon_{1,4} & \varepsilon_{2,4} \\ \varepsilon_{1,5} & \varepsilon_{2,5} \\ \hline \varepsilon_{1,6} & \varepsilon_{2,6} \\ \varepsilon_{1,7} & \varepsilon_{2,7} \end{pmatrix} \tag{4.11.56}$$

Consider the first block in (4.11.56). Developing the first entry of row 1 yields:

$$0 = c_{11}/\pi_1\pi_4 + \varepsilon_{1,1} \quad \Leftrightarrow \quad c_{11} = -\pi_1\ \pi_4\ \varepsilon_{1,1} \tag{4.11.57}$$

Noting that $\varepsilon_{1,1} \sim N(0,\sigma_{11})$, one concludes:

$$\mathbb{E}(c_{11}) = 0 \qquad Var(c_{11}) = \sigma_{11}(\pi_1\ \pi_4)^2 \tag{4.11.58}$$

Similarly, consider the second block and develop the first entry of row 2 to obtain:

$$\frac{\delta_1\sqrt{s_1}}{\pi_1} = \frac{\sqrt{s_1}1^{\pi_3}}{\pi_1}a_{11}^1 + \varepsilon_{1,2} \quad \Leftrightarrow \quad a_{11}^1 = \frac{\delta_1}{1^{\pi_3}} - \frac{\pi_1}{\sqrt{s_1}1^{\pi_3}}\varepsilon_{1,2} \tag{4.11.59}$$

And from this one obtains:

$$\mathbb{E}(a_{11}^1) = \delta_1 \qquad Var(a_{11}^1) = \left(\frac{\sigma_{11}}{s_1}\right)\pi_1^2 \tag{4.11.60}$$

Proceeding the same way with the other entries of blocks 1 and 2, it is straightforward to recover the full prior mean and variance of the normal-Wishart prior for β, as defined in (4.11.16) and (4.11.28). For Σ, consider the third block and develop the first entry of row 6 to obtain:

$$\sqrt{s_1} = \varepsilon_{1,6} \qquad \Longrightarrow \qquad \mathbb{E}(\varepsilon_{1,6}) = 0 \qquad Var(\varepsilon_{1,6}) = s_1 \tag{4.11.61}$$

Continuing in a similar fashion with the other entries of block 3, one recovers the normal-Wishart prior for Σ implied by (4.11.30).

To implement the dummy observation prior, we augment the actual data with the artificial observations. Specifically, we define:

$$Y_d = \begin{pmatrix} Y_{dum} \\ Y \end{pmatrix} \qquad X_d = \begin{pmatrix} X_{dum} \\ X \end{pmatrix} \qquad T_d = T + T_{dum} \tag{4.11.62}$$

The model is then estimated using the posteriors (4.11.51) and (4.11.52), except that the posterior parameters (4.11.50) and (4.11.53) are computed with Y_d, X_d and T_d in place of Y, X and T .

11.6 A large Bayesian VAR prior

The normal-Wishart and dummy observation priors can handle large Bayesian VARs efficiently. They suffer however from two main limits: the prior dependence of β on Σ, which represents a strong and often unrealistic assumption; and the specific Kronecker structure of the prior, which forces to set the cross-variable shrinkage parameter π_2 in the Minnesota prior to 1, another undesirable assumption.

This final section introduces a model that circumvents these two issues and settles an efficient estimation procedure for large Bayesian VARs while preserving a fully flexible and independent Minnesota prior. The trick consists in estimating the model equation by equation rather than all at once, which permits again to deal with smaller matrix inversions. Unlike the normal-Wishart and dummy observation priors however, no analytical solutions are available in this case and simulation methods must be used.

Consider again the general VAR model (4.11.1) written in compact form (4.11.2). Because the model is going to be estimated equation by equation, it is useful to notice that equation i of the VAR model can obtain from column i in (4.11.2) in the following way:

$$Y_i = X\beta_i + \mathcal{E}_i \qquad i = 1,\cdots,n \tag{4.11.63}$$

where Y_i, β_i and $\mathcal{E}_i$ represent column i of $Y, \mathcal{B}$ and $\mathcal{E}$ such that:

$$Y = \begin{pmatrix} Y_i \cdots Y_n \end{pmatrix} \qquad Y_i = \begin{pmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,T} \end{pmatrix} \qquad \mathcal{B} = \begin{pmatrix} \beta_i \cdots \beta_n \end{pmatrix} \qquad \mathcal{E} = \begin{pmatrix} \mathcal{E}_i \cdots \mathcal{E}_n \end{pmatrix} \qquad \mathcal{E}_i = \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,T} \end{pmatrix} \tag{4.11.64}$$

The residual variance-covariance matrix Σ can also be rewritten in equation-specific form by applying a triangular decomposition (property m.30):

$$\Phi \, \Sigma \, \Phi' = \Lambda \qquad \Leftrightarrow \qquad \Sigma = \Phi^{-1} \, \Lambda \, \Phi^{-1\prime} \tag{4.11.65}$$

Λ is a diagonal matrix with positive entries only, while Φ^{-1} (and Φ, by property m.31) are unit lower triangular matrices. Specifically:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} \qquad \Phi = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \phi_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \phi_{n1} & \cdots & \phi_{n(n-1)} & 1 \end{pmatrix} \tag{4.11.66}$$

Λ represents the volatility components of Σ, each λ_i being a positive scaling term which governs the residual variance of equation i of the model. On the other hand, Φ can be interpreted as the (inverse) covariance component of Σ. Denoting by ϕ_i the vector of non-zero and non-one terms in row i of Φ so that $\phi_i = (\phi_{i1} \ \cdots \ \phi_{i(i-1)})'$, ϕ_i then represents the covariance between the residual of equation i of the model and the other shocks.

The parameters of interest of the model are thus the series of equation-specific VAR coefficients $\beta_1, \cdots, \beta_n$, the equation-specific residual variance terms $\lambda_1, \cdots, \lambda_n$, and the equation-specific covariance terms $\phi_2, \cdots, \phi_n$. Assuming independence between the β_i's, λ_i's and ϕ_i's, Bayes rule obtains directly as:

$$\pi(\beta, \lambda, \phi | y) \propto f(y|\beta, \lambda, \phi) \left(\prod_{i=1}^{n} \pi(\beta_i) \right) \left(\prod_{i=1}^{n} \pi(\lambda_i) \right) \left(\prod_{i=2}^{n} \pi(\phi_i) \right) \tag{4.11.67}$$

The likelihood function for the model is still given by (4.11.6). This is however a joint formulation that cannot be exploited to estimate the model on an equation-per-equation basis. Hence, a reformulation is required to express the likelihood function in terms of separate elements β_i, λ_i, and ϕ_i. This is no easy task, but after some work the likelihood function can be shown (book 2, p. 61) to rewrite as:

$$f(y|\beta, \lambda, \phi) = (2\pi)^{-nT/2} \left(\prod_{i=1}^{n} \lambda_i^{-T/2} \right) exp \left(-\frac{1}{2} \sum_{i=1}^{n} \lambda_i^{-1} (Y_i - X\beta_i + \mathcal{E}_{-i} \phi_i)' (Y_i - X\beta_i + \mathcal{E}_{-i} \phi_i) \right) \tag{4.11.68}$$

with $\mathcal{E}_{-i}$ the $T \times (i-1)$ matrix defined as:

$$\mathcal{E}_{-i} = \begin{pmatrix} \mathcal{E}_1 & \mathcal{E}_2 & \cdots & \mathcal{E}_{i-1} \end{pmatrix} \tag{4.11.69}$$

Consider now the priors for β_i, $i = 1, \cdots, n$. The usual Minnesota prior described by equations (4.11.16) - (4.11.20) defines the prior mean b and prior variance and V for all equations simultaneously. Here we use the same prior but simply split b and V into their equation-specific components $b_1, b_2, \cdots, b_n$ and $V_1, V_2, \cdots, V_n$. Each b_i is thus a $k \times 1$ vector of prior mean, while V_i is a $k \times k$ matrix of prior variance. Following, the prior for each β_i is multivariate normal distribution with mean b_i and variance V_i: $\pi(\beta_i) \sim N(b_i, V_i)$. Hence:

$$\pi(\beta_i) = (2\pi)^{-k/2} |V_i|^{-1/2} \exp \left(-\frac{1}{2} (\beta_i - b_i)' V_i^{-1} (\beta_i - b_i) \right) \qquad i = 1, \cdots, n \tag{4.11.70}$$

Since the λ_i parameters are positive terms, an inverse gamma prior is suitable. The priors are thus defined as inverse gamma distributions with shape $\alpha/2$ and scale $\psi/2$: $\pi(\lambda_i) \sim IG(\alpha/2, \psi/2)$. Following:

$$\pi(\lambda_i) = \frac{(\psi/2)^{\alpha/2}}{\Gamma(\alpha/2)} \lambda_i^{-\alpha/2-1} \exp\left(-\frac{\psi}{2\lambda_i}\right) \qquad i = 1, \cdots, n \tag{4.11.71}$$

We typically want to set a weakly informative prior about λ_i by setting small for the prior shape and scales such as $\alpha = \psi = 0.0001$.

Finally, for the covariance vectors ϕ_i we use a weakly informative multivariate normal distribution that leaves the estimation burden to the data: $\pi(\phi_i) \sim N(0, \tau I_{i-1})$, with τ some large value such as $\tau = 1000$. Following:

$$\pi(\phi_i) = (2\pi)^{-(i-1)/2}\ \tau^{-(i-1)/2}\ \exp\left(-\frac{1}{2}\tau^{-1}\phi_i'\phi_i\right) \qquad i = 1, \cdots, n \tag{4.11.72}$$

Substituting for (4.11.68), (4.11.70), (4.11.71) and (4.11.72) in Bayes rule (4.11.67) yields:

$$\begin{aligned}
\pi(\beta, \lambda, \phi|y) \ \propto & \left(\prod_{i=1}^{n} \lambda_i^{-T/2}\right) exp\left(-\frac{1}{2}\sum_{i=1}^{n} \lambda_i^{-1}(Y_i - X\beta_i + \mathcal{E}_{-i}\phi_i)'(Y_i - X\beta_i + \mathcal{E}_{-i}\phi_i)\right) \\
& \times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\beta_i - b_i)'V_i^{-1}(\beta_i - b_i)\right) \\
& \times \prod_{i=1}^{n} \lambda_i^{-\alpha/2-1} \exp\left(-\frac{\psi}{2\lambda_i}\right) \\
& \times \prod_{i=2}^{n} \exp\left(-\frac{1}{2}\tau^{-1}\phi_i'\phi_i\right)
\end{aligned} \tag{4.11.73}$$

where as usual any multiplicative term not involving β_i, λ_i or γ_i has been relegated to the normalization constant. This joint posterior cannot be integrated out to obtain the marginal posteriors, so the Gibbs sampling algorithm must be used.

Obtain first the conditional posterior $\pi(\beta_i|y, \beta_{-i})$[3]. From definition 6.1, this is done by starting from the joint posterior (4.11.73) and relegating to the normalization constant any multiplicative term not involving β_i, yielding:

$$\pi(\beta_i|y, \beta_{-i}) \propto exp\left(-\frac{1}{2}\lambda_i^{-1}(Y_i - X\beta_i + \mathcal{E}_{-i}\phi_i)'(Y_i - X\beta_i + \mathcal{E}_{-i}\phi_i)\right) \times \exp\left(-\frac{1}{2}(\beta_i - b_i)'V_i^{-1}(\beta_i - b_i)\right) \tag{4.11.74}$$

Rearrange and complete the squares (book 2, p. 64) to obtain:

$$\pi(\beta_i|y, \beta_{-i}) \propto exp\left(-\frac{1}{2}(\beta_i - \bar{b}_i)'\bar{V}_i^{-1}(\beta_i - \bar{b}_i)\right) \tag{4.11.75}$$

with:

$$\bar{V}_i = (V_i^{-1} + \lambda_i^{-1}X'X)^{-1} \qquad \bar{b}_i = \bar{V}_i(V_i^{-1}b_i + \lambda_i^{-1}X'[Y_i + \mathcal{E}_{-i}\phi_i]) \tag{4.11.76}$$

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

[3]For any parameter θ_i we use $\pi(\theta_i|\theta_{-i})$ to denote the density of θ_i conditional on all model parameters except θ_i.

(4.11.76) makes it clear why this prior is efficient: the matrix $\bar{V}_i$ is only of dimension $k \times k$, reducing the inversion burden by a factor n^2 compared to the independent prior where $\bar{V}$ is of dimension $kn \times kn$[4].

Obtain then the conditional posterior $\pi(\lambda_i|y,\lambda_{-i})$. Start from (4.11.73) and relegate to the normalization constant any multiplicative term not involving λ_i to obtain:

$$\pi(\lambda_i|y,\lambda_{-i}) \propto \lambda_i^{-T/2} \times exp\left(-\frac{1}{2}\lambda_i^{-1}(Y_i - X\beta_i + \mathcal{E}_{-i}\phi_i)'(Y_i - X\beta_i + \mathcal{E}_{-i}\phi_i)\right) \times \lambda_i^{-\alpha/2-1} \exp\left(-\frac{\psi}{2\lambda_i}\right) \tag{4.11.77}$$

And this immediately rewrites as:

$$\pi(\lambda_i|y,\lambda_{-i}) \propto \lambda_i^{-\bar{\alpha}/2-1}\, exp\left(-\frac{\bar{\psi}_i}{2\lambda_i}\right) \tag{4.11.78}$$

with:

$$\bar{\alpha} = \alpha + T \qquad\qquad \bar{\psi}_i = \psi + (\mathcal{E}_i + \mathcal{E}_{-i}\phi_i)'(\mathcal{E}_i + \mathcal{E}_{-i}\phi_i) \tag{4.11.79}$$

This is the kernel of an inverse gamma distribution with shape $\bar{\alpha}/2$ and scale $\bar{\psi}_i/2$: $\pi(\lambda_i|y,\lambda_{-i}) \sim IG(\bar{\alpha}/2, \bar{\psi}_i/2)$.

Obtain finally the conditional posterior $\pi(\phi_i|y,\phi_{-i})$. Start from (4.11.73) and relegate to the normalization constant any multiplicative term not involving ϕ_i to obtain:

$$\pi(\phi_i|y,\phi_{-i}) \propto exp\left(-\frac{1}{2}\lambda_i^{-1}(Y_i - X\beta_i + \mathcal{E}_{-i}\phi_i)'(Y_i - X\beta_i + \mathcal{E}_{-i}\phi_i)\right) \times \exp\left(-\frac{1}{2}\tau^{-1}\phi_i'\phi_i\right) \tag{4.11.80}$$

After rearranging and completing the squares (book 2, p. 66), this becomes:

$$\pi(\phi_i|y,\phi_{-i}) \propto exp\left(-\frac{1}{2}(\phi_i - \bar{f}_i)'\bar{Z}_i^{-1}(\phi_i - \bar{f}_i)\right) \tag{4.11.81}$$

with:

$$\bar{Z}_i = (\tau^{-1}I_{i-1} + \lambda_i^{-1}\mathcal{E}'_{-i}\mathcal{E}_{-i})^{-1} \qquad\qquad \bar{f}_i = \bar{Z}_i(-\lambda_i^{-1}\mathcal{E}'_{-i}\mathcal{E}_i) \tag{4.11.82}$$

This is the kernel of a multivariate normal distribution with mean $\bar{f}_i$ and variance $\bar{Z}_i$: $\pi(\phi_i|y,\phi_{-i}) = N(\bar{f}_i, \bar{Z}_i)$.

[4]Precisely, the number of flops (basic operations) to invert a $n \times n$ matrix is of the order of $\mathcal{O}(n^3)$. Thus inverting the n matrices $\bar{V}_i$ of dimension $k \times k$ requires $n \times k^3$ flops, while inverting the single $nk \times nk$ matrix $\bar{V}$ from the independent prior takes n^3k^3 flops, that is, n^2 times more operations.

We can now introduce the Gibbs sampling algorithm for the large Bayesian VAR prior.

algorithm 11.2: Gibbs sampling algorithm for the large Bayesian VAR prior

1. set initial values $\beta_i^{(0)}$, $\lambda_i^{(0)}$ and $\phi_i^{(0)}$, $i = 1, \cdots, n$. We use maximum likelihood estimates: $\beta_i^{(0)} = \hat{\beta}_i$, $\lambda_i^{(0)} = \hat{\Sigma}_{ii}$, and for $\phi_i^{(0)}$ we use the covariance entries of row i of $\hat{\Sigma}^{-1}$.

2. at iteration j, for $i = 1, \cdots, n$, draw:

   $\beta_i^{(j)}$ from $\pi(\beta_i | y, \beta_{-i}) = N(\bar{b}_i, \bar{V}_i)$ with:

   $\bar{V}_i = (V_i^{-1} + \lambda_i^{-1} X'X)^{-1} \qquad \bar{b}_i = \bar{V}_i (V_i^{-1} b_i + \lambda_i^{-1} X'[Y_i + \mathcal{E}_{-i} \phi_i])$

   Update $\mathcal{E}_i$ and $\mathcal{E}_{-i}$.

3. at iteration j, for $i = 1, \cdots, n$, draw:

   $\lambda_i^{(j)}$ from $\pi(\lambda_i | y, \lambda_{-i}) \sim IG(\bar{\alpha}/2, \bar{\psi}_i/2)$ with:

   $\bar{\alpha} = \alpha + T \qquad \bar{\psi}_i = \psi + (\mathcal{E}_i + \mathcal{E}_{-i} \phi_i)' (\mathcal{E}_i + \mathcal{E}_{-i} \phi_i)$

   Update Λ.

4. at iteration j, for $i = 2, \cdots, n$, draw:

   $\phi_i^{(j)}$ from $\pi(\phi_i | y, \phi_{-i}) = N(\bar{f}_i, \bar{Z}_i)$ with:

   $\bar{Z}_i = (\tau^{-1} I_{i-1} + \lambda_i^{-1} \mathcal{E}'_{-i} \mathcal{E}_{-i})^{-1} \qquad \bar{f}_i = \bar{Z}_i (-\lambda_i^{-1} \mathcal{E}'_{-i} \mathcal{E}_i)$

   Update Φ.

5. at iteration j:

   Obtain $\beta^{(j)}$ from $\beta^{(j)} = vec(\mathcal{B}^{(j)})$, with $\mathcal{B}^{(j)} = (\beta_1^{(j)} \; \beta_2^{(j)} \; \cdots \; \beta_n^{(j)})$

   Calculate Φ^{-1}, then recover $\Sigma^{(j)}$ from $\Sigma^{(j)} = \Phi^{-1} \, \Lambda \, \Phi^{-1\prime}$.

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

CHAPTER 12

Further aspects of Bayesian vector autoregressions

12.1 Constrained coefficients

Sometimes, we want to constrain certain coefficients of the VAR to take predefined values. This typically happens when economic theory provides a rationale for the value of some coefficients or equations of the model. A classical example is the case of a VAR model that involves a large economy like the United States, and a small open economy like Jamaica. Jamaica is certainly impacted by economic activity in the United States (more than 40% of Jamaican exports go to the US), but the converse is not true: Jamaica has most likely no impact on overall eonomic activity in the United States.

To represent the situation, we set a simple 2-variable VAR model with p lags and one constant, where $y_{1,t}$ is GDP growth for the United States and $y_{2,t}$ is GDP growth for Jamaica:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} c_{11} \\ c_{21} \end{pmatrix} + \begin{pmatrix} a_{11}^1 & a_{12}^1 \\ a_{21}^1 & a_{22}^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \cdots + \begin{pmatrix} a_{11}^p & a_{12}^p \\ a_{21}^p & a_{22}^p \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ y_{2,t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \tag{4.12.1}$$

Assuming that Jamaican GDP growth has no effect on GDP growth in the United States is equivalent to imposing the constraint $a_{12}^1 = \cdots = a_{12}^p = 0$, so that none of the lags of Jamaican GDP growth ever affect GDP growth in the United States. This is the main principle behind **Granger causality**, which determines if some variable contributes overall in predicting another variable of the model.

Granger causality is just one possible application of constrained coefficients. The concept is in fact general and can apply to any of the VAR coefficients, endogenous or exogenous. Consider the vector β of VAR coefficients defined in (4.11.5), and assume we want to constrain coefficient β_k to take value $\tilde{b}_k$. To do so, we simply take the prior parameters b and V of the Minnesota prior and replace the corresponding entries b_k and V_{kk} with $\tilde{b}_k$ and $\tilde{V}_{kk}$. By setting $\tilde{V}_{kk}$ to some arbitrarily small value such as $1e^{-10}$, the posterior distribution can be made arbitrarily tight around $\tilde{b}_k$, effectively turning β_k into an almost constant coefficient. We may also impose a softer constraint by setting $\tilde{V}_{kk}$ to a moderately small value like 0.1, which centers the posterior around $\tilde{b}_k$ but allows for some variability.

Note finally that constrained coefficients can only be used with the Minnesota, independent and large BVAR priors. The specific prior structure of the normal-Wishart and dummy observation priors implies that the variance of one equation is proportional to the variance of the other equations, so that the variance of the constrained coefficients would be replicated across all equations. This is an undesired effect that would produce meaningless posterior estimates.

12.2 Dummy observation extensions

Dummy observations have been introduced in section 11.5. This section shows how to extend the concept to any prior distribution, following the same logic as before: create artifical observations that will supplement the data with additional prior information through the likelihood function. We develop three classical dummy observation extensions: the **sums of coefficients** approach of Doan et al. (1984); the **dummy initial observation** introduced by Sims (1993); and the **long run prior** of Giannone et al. (2019).

These three extnesions are closely related to the concept of **cointegration**. They are mostly used whenever one estimates a VAR that includes economic variables in level. Such variables are typically $I(1)$, involving a non-stationary behaviour of the VAR. If instead there exists a cointegration relation between the variables, the model will be stationary even each variable included in it is effectively $I(1)$. The sums of coefficients extension implements the belief of a unit root in the model; the dummy initial observation drives the model towards cointegration; the long-run prior aims at discriminating between stationary and nonstationary linear combinations, defining their priors accordingly.

Consider again the general VAR model (4.11.1), rewritten here for convenience:

$$y_t = Cz_t + A_1 y_{t-1} + \cdots + A_p y_{t-p} + \varepsilon_t \tag{4.12.2}$$

It can be shown (book 2, p. 67) that this model can rewrite in **error correction form** as:

$$\Delta y_t = Cz_t + (A_1 + \cdots + A_p - I)y_{t-1} + B_1 \Delta y_{t-1} + \cdots + B_{p-1}\Delta y_{t-(p-1)} + \varepsilon_t \qquad B_i \equiv -\sum_{j=i+1}^{p} A_j \tag{4.12.3}$$

Looking at the right-hand side of (4.12.3), we see the stationary terms $Cz_t, B_1\Delta y_{t-1}, \cdots, B_{p-1}\Delta y_{t-(p-1)}$, and the possibly non-stationary **error correction term** $(A_1 + \cdots + A_p - I)y_{t-1}$. If the latter is $I(0)$, then there exists at least one cointegration relation and the model is stationary. If instead we have:

$$A_1 + \cdots + A_p - I = 0 \tag{4.12.4}$$

Then the reformulated VAR model (4.12.3) becomes:

$$y_t = y_{t-1} + Cz_t + B_1 \Delta y_{t-1} + \cdots + B_{p-1}\Delta y_{t-(p-1)} + \varepsilon_t \tag{4.12.5}$$

In this case, the model has a unit root and cointegration is ruled out. The **sums-of-coefficients** extension then consists in implementing the prior belief (4.12.4) by the way of dummy observations. Specifically, we create the following dummy observations:

$$Y_{sum} = diag(\bar{y}_1/\pi_5 \ \cdots \ \bar{y}_n/\pi_5) \qquad X_{sum} = (0_{n\times m} \quad \mathbf{1}_p' \otimes Y_{sum}) \tag{4.12.6}$$

where $\bar{y}_i$ denotes the arithmetic mean of each endogenous variable in the VAR (possibly calculated from an initial sample of data), and π_5 is a shrinkage hyperparameter specific to the sums of coefficient extension. To see how this relates to (4.12.4), consider again a simple VAR model with 2 variables, 2 lags and a constant. Using again the compact VAR formulation (4.11.2) with the dummy observations (4.12.6) yields:

$$\begin{pmatrix} \bar{y}_1/\pi_5 & 0 \\ 0 & \bar{y}_2/\pi_5 \end{pmatrix} = \begin{pmatrix} 0 & \bar{y}_1/\pi_5 & 0 & \bar{y}_1/\pi_5 & 0 \\ 0 & 0 & \bar{y}_2/\pi_5 & 0 & \bar{y}_2/\pi_5 \end{pmatrix} \begin{pmatrix} c_{11} & c_{21} \\ a^1_{11} & a^1_{21} \\ a^1_{12} & a^1_{22} \\ a^2_{11} & a^2_{21} \\ a^2_{12} & a^2_{22} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{1,2} \\ \varepsilon_{2,1} & \varepsilon_{2,2} \end{pmatrix} \tag{4.12.7}$$

Develop the top-left entry to obtain:

$$\bar{y}_1/\pi_5 = \bar{y}_1/\pi_5 a_{11}^1 + \bar{y}_1/\pi_5 a_{11}^2 + \varepsilon_{1,1} \Rightarrow a_{11}^1 + a_{11}^2 - 1 = -\frac{\pi_5}{\bar{y}_1}\varepsilon_{1,1} \tag{4.12.8}$$

And from this one obtains:

$$\mathbb{E}(a_{11}^1 + a_{11}^2 - 1) = 0 \qquad Var(a_{11}^1 + a_{11}^2 - 1) = \left(\frac{\pi_5}{\bar{y}_1}\right)^2 \sigma_{11} \tag{4.12.9}$$

This replicates the top left entry of (4.12.4). Repeating the procedure with the other dummy observations replicates (4.12.4) as a whole. The effect of the shrinkage parameter π_5 is clear from (4.12.9): the smaller π_5, the smaller the prior variance on the sums of coefficients, with $\pi_5 \to \infty$ implying a diffuse prior while $\pi_5 \to 0$ forces a unit root in each equation.

A major limit of the sums of coefficient approach is that it eliminates the possibility of cointegration. To remedy this issue, the **dummy initial observation** extension was developed. It consists in a single artificial observation, defined as:

$$Y_{obs} = (\bar{y}_1/\pi_6 \ \cdots \ \bar{y}_n/\pi_6) \qquad X_{obs} = (\bar{z}/\pi_6 \quad \mathbf{1}_p' \otimes Y_{obs}) \tag{4.12.10}$$

where $\bar{z}$ denotes the arithmetic mean of the exogenous variables of the VAR and π_6 is a shrinkage hyper-parameter specific to the dummy initial observation extension. Considering the simple VAR model with 2-variables, 2 lags and a constant, using again the compact VAR formulation (4.11.2) with the dummy observations (4.12.10) yields:

$$\begin{pmatrix}\bar{y}_1/\pi_6 & \bar{y}_2/\pi_6\end{pmatrix} = \begin{pmatrix}1/\pi_6 & \bar{y}_1/\pi_6 & \bar{y}_2/\pi_6 & \bar{y}_1/\pi_6 & \bar{y}_2/\pi_6\end{pmatrix} \begin{pmatrix} c_{11} & c_{21} \\ a_{11}^1 & a_{21}^1 \\ a_{12}^1 & a_{22}^1 \\ a_{11}^2 & a_{21}^2 \\ a_{12}^2 & a_{22}^2 \end{pmatrix} + \begin{pmatrix}\varepsilon_{1,1} & \varepsilon_{1,2}\end{pmatrix} \tag{4.12.11}$$

Develop the first entry to obtain:

$$\bar{y}_1/\pi_6 = c_{11}/\pi_6 + a_{11}^1\bar{y}_1/\pi_6 + a_{12}^1\bar{y}_2/\pi_6 + a_{11}^2\bar{y}_1/\pi_6 + a_{12}^2\bar{y}_2/\pi_6 + \varepsilon_{1,1} \tag{4.12.12}$$

And this rewrites as:

$$\bar{y}_1 = c_{11} + a_{11}^1\bar{y}_1 + a_{12}^1\bar{y}_2 + a_{11}^2\bar{y}_1 + a_{12}^2\bar{y}_2 + \pi_6\varepsilon_{1,1} \tag{4.12.13}$$

Finally, calculating expectation and variance yields:

$$\bar{y}_1 = c_{11} + a_{11}^1\bar{y}_1 + a_{12}^1\bar{y}_2 + a_{11}^2\bar{y}_1 + a_{12}^2\bar{y}_2 \qquad Var(\bar{y}_1 - c_{11} - a_{11}^1\bar{y}_1 - a_{12}^1\bar{y}_2 - a_{11}^2\bar{y}_1 - a_{12}^2\bar{y}_2) = \pi_6^2\sigma_{11} \tag{4.12.14}$$

This representation states that a no-change forecast for all variables is a good forecast for the observed sample. In this case, either all the variables are at their unconditional mean, or the system is characterized by the presence of an unspecified number of unit roots and the variables share a common stochastic trend. The dummy initial observation prior is then consistent with the existence of cointegration. The shrinkage parameter π_6 controls the tightness of the prior, the left-hand-side of (4.12.14) holding exactly when $\pi_6 \to 0$.

The dummy initial observation makes the model consistent with cointegration but is agnostic about the form that cointegration may take. The **long run prior** constitutes an alternative that explicitely models the possible forms of cointegration as well as other possible linear relations between the model variables. This is done by the way of some matrix J that identifies the relevant linear combinations that may exist between the variables entering the model, where J is squared and invertible. Different degrees of shrinkage

are then applied to the different combinations to either favor a unit root (tight shrinkage around a sums of coefficient equal to 0) or permit cointegration (loose shrinkage to allow for some cointegration relation).

To make things more concrete, consider again a simple VAR model with 2 variables. Let us be specific for this example[1] and assume these two variables are the log of output ($y_{1,t} = log(Y_t)$) and the log of real investment ($y_{2,t} = log(I_t)$). Economic theory suggests that output and investment are likely to share a common trend (so $y_{1,t} + y_{2,t}$ is $I(1)$), while the log investment-to-output ratio is stationary (so $y_{2,t} - y_{1,t}$ is $I(0)$). This produces the following J matrix:

$$J = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \tag{4.12.15}$$

Once J is defined, the following dummy observations are created:

$$Y_{lrp} = diag(J\,\bar{y}/\pi_7)\,J^{-1\prime} \qquad X_{lrp} = (0_{n\times m} \quad \mathbf{1}'_p \otimes Y_{lrp}) \tag{4.12.16}$$

where $\bar{y}$ denotes the vector of arithmetic means of the VAR variables (possibly calculated from an initial sample of data), and π_7 is a shrinkage hyperparameter specific to the long run prior extension. Using the compact VAR formulation (4.11.2), it can be shown (book 2, p. 67) that the dummy observations imply:

$$(A_1 + \cdots + A_p - I)J^{-1} = -\,diag(\pi_7/J\bar{y})\;\; \mathcal{E}'_{lrp} \tag{4.12.17}$$

Taking expectations and variance from (4.12.17) directly yields:

$$\mathbb{E}\left[(A_1 + \cdots + A_p - I)J^{-1}\right] = 0 \qquad Var\left[vec\{(A_1 + \cdots + A_p - I)J^{-1}\}\right] = diag(\pi_7^2/(J\bar{y})^2) \otimes \Sigma \tag{4.12.18}$$

The matrix $(A_1 + \cdots + A_p - I)J^{-1}$ captures the effect of the linear combinations defined in J on Δy_t, effectively setting a prior on the error correction term of the VAR. For instance, the sums of coefficients extension can be seen as a special case of the long run prior with $J = I$, defining the prior belief that each variable of the model behaves as an individual random walk. The shrinkage parameter π_7 controls the tightness of the prior on the linear combinations J, along with the additional shrinkage term $J\bar{y}$. As the absolute value of $J_i\bar{y}$ (the linear combination defined by row i of J) is typically smaller for mean-reverting combinations of y_t, the prior shrinkage will be less for cointegrating relations than for other non-stationary combinations. This mechanism automatically generates softer constraints on cointegration relations while still maintaining the balance between stationary relations and non-stationary dynamics.

12.3 Marginal likelihood

The marginal likelihood constitutes a key element of VAR modelling as it constitutes the basis of model comparison and hypothesis testing. The marginal likelihood cannot be computed for all VAR models, but can be estimated for the Minnesota, normal-Wishart and independent priors.

Consider first the Minnesota prior developed in section 11.2. From definition 4.6 of the marginal likelihood, the likelihood function (4.11.6) and the prior (4.11.12), one obtains:

$$\begin{aligned} f(y) = \int f(y|\beta)\pi(\beta)d\beta \;\; &= \int (2\pi)^{-nT/2}|\bar{\Sigma}|^{-1/2} exp\left(-\frac{1}{2}(y - \bar{X}\beta)'\bar{\Sigma}^{-1}(y - \bar{X}\beta)\right) \\ &\quad \times (2\pi)^{-q/2}|V|^{-1/2}\exp\left(-\frac{1}{2}(\beta - b)'V^{-1}(\beta - b)\right)d\beta \end{aligned} \tag{4.12.19}$$

[1] This exemple is provided in Giannone et al. (2019). Researchers interested in finding suggestions of relevant restrictions for other economic variables may also consult section 5 of the paper which provides an example with a larger model.

Rearranging and completing the squares, this reformulates as (book 2, p. 68):

$$\begin{aligned} f(y) &= (2\pi)^{-nT/2}|\Sigma|^{-T/2}|I_q + V(\Sigma^{-1}\otimes X'X)|^{-1/2} exp\left(-\frac{1}{2}\left[b'V^{-1}b - \bar{b}'\bar{V}^{-1}\bar{b} + tr(Y'Y\Sigma^{-1})\right]\right) \\ &\times \int (2\pi)^{-q/2}\bar{V}^{-1} exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right) d\beta \end{aligned} \tag{4.12.20}$$

The first row of (4.12.20) can get out of the integral as it does not involve β. The second row is recognised as the probability density function of a multivariate normal distribution which thus integrates to 1, eventually leaving:

$$f(y) = (2\pi)^{-nT/2}|\Sigma|^{-T/2}|I_q + V(\Sigma^{-1}\otimes X'X)|^{-1/2} exp\left(-\frac{1}{2}\left[b'V^{-1}b - \bar{b}'\bar{V}^{-1}\bar{b} + tr(Y'Y\Sigma^{-1})\right]\right) \tag{4.12.21}$$

Consider now the marginal likelihood for the normal-Wishart prior introduced in section 11.3. From definition 4.6 of the marginal likelihood, the likelihood function (4.11.23) and the priors (4.11.24) and (4.11.29), one obtains:

$$\begin{aligned} f(y) &= \int\int (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left(-\frac{1}{2}(\beta-\hat{\beta})'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}(\beta-\hat{\beta})\right) \\ &\times exp\left(-\frac{1}{2}\,tr\left[\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right]\right) \\ &\times (2\pi)^{-q/2}|\Sigma\otimes W|^{-1/2}\exp\left(-\frac{1}{2}(\beta-b)'(\Sigma\otimes W)^{-1}(\beta-b)\right) \\ &\times \frac{2^{-\alpha n/2}}{\Gamma_n\left(\frac{\alpha}{2}\right)}|S|^{\alpha/2}|\Sigma|^{-(\alpha+n+1)/2} exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}S\right\}\right) d\beta d\Sigma \end{aligned} \tag{4.12.22}$$

Rearranging and completing the squares, this reformulates as (book 2, p. 69):

$$\begin{aligned} f(y) &= \pi^{-nT/2}\ |I_k + WX'X|^{-n/2}\ |S|^{-T/2}|I_n + S^{-1}(\bar{S}-S)|^{-\bar{\alpha}/2}\ \frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha}{2}\right)} \\ &\times \int\int (2\pi)^{-nk/2}|\Sigma|^{-k/2}|\bar{W}|^{-n/2} exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}(\mathcal{B}'-\bar{B})'\bar{W}^{-1}(\mathcal{B}'-\bar{B})\right\}\right) d\beta \\ &\times \frac{2^{-\bar{\alpha}n/2}}{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}|\bar{S}|^{\bar{\alpha}/2}|\Sigma|^{-(\bar{\alpha}+n+1)/2} exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}\bar{S}\right\}\right) d\Sigma \end{aligned} \tag{4.12.23}$$

where $\bar{B}, \bar{W}, \bar{S}$ and $\bar{\alpha}$ are defined as in (4.11.33). The inner integral with respect to β in the second row is recognized as the density function of a matrix normal distribution which therefore integrates to 1 and simplifies out. It then remains the outer integral with respect to Σ in the third row which is recognized as the density function of an inverse Wishart distribution which also integrates to 1. What ultimately remains of (4.12.23) after integration is thus simply:

$$f(y) = \pi^{-nT/2}\ |I_k + WX'X|^{-n/2}\ |S|^{-T/2}|I_n + S^{-1}(\bar{S}-S)|^{-\bar{\alpha}/2}\ \frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha}{2}\right)} \tag{4.12.24}$$

Consider finally the independent prior developed in section 11.4. The model relies on simulation methods, so the marginal likelihood must be computed from equation (2.6.15), namely:

$$f(y) \approx \frac{f(y|\beta^*,\Sigma^*)\pi(\beta^*,\Sigma^*)}{\pi(\Sigma^*|y,\beta^*) \times \frac{1}{J}\sum_{j=1}^{J}\pi(\beta^*|\Sigma^{(j)},y)} \tag{4.12.25}$$

Using the likelihood function (4.11.6), the priors (4.11.12) and (4.11.29), and the conditional posteriors (4.11.42) and (4.11.45), it can be shown that the marginal likelihood formulates as (book 2, p. 72):

$$
\begin{aligned}
f(y) \;&\approx \pi^{-nT/2}\;|S|^{-T/2}|I_n + S^{-1}(\bar{S}-S)|^{-\bar{\alpha}/2}\;\frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha}{2}\right)} \\
&\times \frac{\exp\left(-\frac{1}{2}(\beta-b)'V^{-1}(\beta-b)\right)}{\frac{1}{J}\sum_{j=1}^{J}|I_q + V(\Sigma^{-1}\otimes X'X)|^{1/2}\exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{V}^{-1}(\beta-\bar{b})\right)}
\end{aligned}
\tag{4.12.26}
$$

This form is similar to (4.12.24), save for the approximation of the determinant term stemming from the Gibbs sampler.

12.4 Stationary priors

This section introduces the notions of stability and stationarity. Indeed, econometricians are often interested in avoiding explosive behaviours for a VAR model, and this can be handled easily within a Bayesian framework.

Consider the general VAR model (4.11.1). For our purpose, it is convenient to rewrite it as a VAR(1) model doing the following:

$$
\begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{pmatrix} = \begin{pmatrix} Cz_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} A_1 & A_2 & \cdots & A_p \\ I_n & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I_n & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4.12.27}
$$

Or, in compact form:

$$
\gamma_t = \mu_t + F\gamma_{t-1} + \zeta_t \tag{4.12.28}
$$

(4.12.28) is known as the **companion form** of the VAR model. Because it is expressed as a simple VAR(1), it is easy to use back recursion to obtain (book 2, p. 74):

$$
\gamma_t = \sum_{i=0}^{j} F^i \mu_{t-i} + F^j \gamma_{t-j} + \sum_{i=0}^{j} F^i \zeta_{t-i} \tag{4.12.29}
$$

We can decompose the dynamic matrix F into $F = Q\Lambda Q'$, where Q and Λ respectively denote the matrix of eigenvectors and eigenvalues of F. From property m.61, if all the eigenvalues of F are smaller than 1 in absolute value then $F^j \to 0$ as $j \to \infty$. Following, the impact of any past shock and exogenous regressor in (4.12.29) eventually dies out, and so does the impact of the initial condition γ_{t-j}. In this case, the VAR model does not display an explosive behaviour. We can define stability formally as:

definition 12.1: a VAR model is **stable** if the eigenvalues of the companion matrix F are all smaller than 1 in absolute value.

In the context of VAR modelling, a closely related concept is that of **stationarity** :

definition 12.2: let y_t be a n-dimensional random vector. y_t is **weakly stationary** if:

$\mathbb{E}(y_t) = \mu$ for all t

$\mathbb{E}(y_t - \mu)(y_{t-i} - \mu) = \Gamma_i$ for all t and all i

In other words, a process is weakly stationary if its first and second moments are invariant to the date t. In this case the process revolves around a constant mean with constant volatility. For a VAR model that includes only a constant as exogenous, we have the following result:

theorem 12.1: a VAR model that is stable is weakly stationary.

Therefore it is sufficient to examine the eigenvalues of the companion matrix F of the VAR to establish weak stationarity. If the VAR is stationary, it has a steady-state value $\mathbb{E}(y_t) = \mu$ from definition 12.2, and this steady-state directly obtains as (book 2, p. 74):

$$\mu = (I - A_1 - \cdots - A_p)^{-1} C\, z_t \tag{4.12.30}$$

If the VAR includes exogenous variables other than a constant, the eigenvalues establish stability but not stationarity. In particular, the mean of the process will be non-constant as it depends on z_t.

Often, econometricians want to constrain the VAR model they estimate to be stationary. This may occur both for technical reasons (avoid explosive behaviours) and theretical reasons (a VAR with stationary data and stationary Minnesota prior is expected to yield a stationary posterior). In a Bayesian context, this means that any value of the dynamic coefficients β obtained from its posterior distribution should result in a stationary VAR model.

It turns out that inducing stationarity is straightforward in a Bayesian context. To do so, one simply adds stationarity as a prior belief in the estimation process. Consider for instance the Minnesota prior developed in section 11.2. The prior distribution for β is $\pi(\beta) \sim N(b, V)$, with density function given by (4.11.12). To reflect the prior belief of stationarity, we implement the same prior but truncate from it any β value that results in non-stationarity. In other words, we replace the multivariate normal prior with a truncated multivariate normal distribution, where the truncation is perfomed on the non-stationary parts of the distribution. This can be done simply by the way of the indicator function $\mathbb{1}(\lambda(F) < 1)$ which takes a value of 1 whenever the eigenvalues $\lambda(F)$ of the companion matrix F are all smaller than 1 in absolute values, and 0 otherwise. Following, the density function of the prior (4.11.12) is replaced with:

$$\pi(\beta) \propto (2\pi)^{-q/2} |V|^{-1/2} \exp\left(-\frac{1}{2} (\beta - b)' V^{-1} (\beta - b) \right) \times \mathbb{1}(\lambda(F) < 1) \tag{4.12.31}$$

This is exactly the same prior as before save for the truncation indicator $\mathbb{1}(\lambda(F) < 1)$. The posterior is thus also the same and given by (4.11.14), save for the additional truncation term inherited from the prior:

$$\pi(\beta|y) \propto \exp\left(-\frac{1}{2} (\beta - \bar{b})' \bar{V}^{-1} (\beta - \bar{b}) \right) \times \mathbb{1}(\lambda(F) < 1) \tag{4.12.32}$$

The truncated posterior ensures that only stationary values of β can be obtained from the posterior distribution. The logic is readily generalized to any other prior: suffice is to truncate the usual prior $\pi(\beta)$ with the indicator function $\mathbb{1}(\lambda(F) < 1)$ to transmit the truncation to the posterior and ensure stationary values. Whenever sampling is involved in the estimation process (e.g. for the independent prior), the truncation is exerted easily by discarding any non-stationary draw obtained from the sampler and drawing new candidates until a stationary value is obtained.

A final word of warning: stationary priors should only be used whenever it is meaningful to do so. For instance, a model that integrates economic variables in level typically implies a non-stationary behaviour. Forcing stationarity with a stationary prior is not only irrelevant from a theoretical point of view, but can also produce undesired behaviours such as reversion of the model to an equilibrium value that does not exist in the true data dynamics.

12.5 Efficient sampling

Surprisingly, even VAR models with analytical posteriors like the Minnesota and the normal-Wishart priors require the simulation of posterior values for β and Σ. This is because all the VAR applications (such as predictions and impulse response functions, to be covered in chapter 13) rely on the Gibbs sampling algorithm, which in turn uses simulated posterior draws for β and Σ. It is thus important to sample efficiently in order to avoid computational bottlenecks during the estimation process.

The independent and large Bayesian VAR priors make heavy use of the multivariate normal distribution in algorithms 11.1 and 11.2. For this reason, the efficient sampling algorithm 9.4 must be used to generate the multivariate normal draws with maximum efficiency.

The normal-Wishart and dummy observation priors produce analytical solutions: a matrix Student posterior for β, and an inverse Wishart distribution for Σ. These analytical distributions are useful to calculate posterior moments like the mean and variance, but sampling from the matrix Student distribution is inefficient as it requires an additional draw from the inverse Wishart distribution. To avoid this, a faster alternative consists in sampling β from its conditional posterior. Indeed, definition 2.12 implies that $\pi(\beta,\Sigma|y) = \pi(\Sigma|y)\pi(\beta|y,\Sigma)$. In other words, to obtain a draw from the joint posterior $\pi(\beta,\Sigma|y)$, one may first sample Σ from the unconditional posterior $\pi(\Sigma|y)$, then sample β from the conditional posterior $\pi(\beta|y,\Sigma)$.

To obtain $\pi(\beta|y,\Sigma)$, we follow definition 6.1 and start from the joint posterior (4.11.31), then relegate to the normalization constant any term not involving β. This yields:

$$\pi(\beta|y,\Sigma) \propto exp\left(-\frac{1}{2}(\beta-\hat{\beta})'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}(\beta-\hat{\beta})\right) \times \exp\left(-\frac{1}{2}(\beta-b)'(\Sigma\otimes W)^{-1}(\beta-b)\right) \tag{4.12.33}$$

After some manipulations (book 2, p. 74):, this rewrites as:

$$\pi(\beta|y,\Sigma) \propto exp\left(-\frac{1}{2}tr\{\Sigma^{-1}(\mathcal{B}-\bar{B})'\bar{W}^{-1}(\mathcal{B}-\bar{B})\}\right) \tag{4.12.34}$$

with $\bar{W}$ and $\bar{B}$ defined as in (4.11.33).This is the kernel of a matrix normal distribution with location $\bar{B}$ and shapes $\bar{W}$ and Σ: $\pi(\beta|y,\Sigma) \sim MN(\bar{B},\bar{W},\Sigma)$. This solution is more efficient as it only requires one single inverse Wishart draw from $\pi(\Sigma|y)$, the generation of matrix normal random numbers involving no inverse Wishart sampling.

CHAPTER 13

Bayesian VAR: basic applications

13.1 Impulse-response function

A major object of interest in VAR modelling is the so-called impulse-response function. Fundamentally, the impulse-response function measures the effect of a unit shock ε_t on the current and future values of the model y_t, y_{t+1}, y_{t+2}. It thus describes the dynamics of the shock transmission over time on the VAR model. To define formally the impulse-response function, we first state a classical result in VAR analysis:

theorem 13.1: (Wold theorem) any weakly stationary VAR model can be expressed as:
$y_t = A(L)^{-1}Cz_t + \Phi_0\varepsilon_t + \Phi_1\varepsilon_{t-1} + \Phi_2\varepsilon_{t-2}\dots$

where $A(L)^{-1}$ denotes the inverse lag polynomial of the VAR coefficients[1]. The Wold theorem fundamentally states that any stationary VAR model can be rewritten as a constant plus an infinite order moving average process. Each Φ_h is a $n \times n$ matrix of coefficients that has the interpretation:

$$\Phi_h = \frac{\partial y_{t+h}}{\partial \varepsilon_t} \qquad \text{with} \qquad \Phi_h = \begin{pmatrix} \phi_{11}^h & \phi_{12}^h & \cdots & \phi_{1n}^h \\ \phi_{21}^h & \phi_{22}^h & \cdots & \phi_{2n}^h \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1}^h & \phi_{n2}^h & \cdots & \phi_{nn}^h \end{pmatrix} \qquad \text{so that} \qquad \phi_{ij}^h = \frac{\partial y_{i,t+h}}{\partial \varepsilon_{j,t}} \tag{4.13.1}$$

Thus, ϕ_{ij}^h identifies the effect of a one-unit increase in $\varepsilon_{j,t}$ on the variable $y_{i,t+h}$, keeping all other innovations at all dates null. The series $\Phi_0, \Phi_1, \Phi_2 \cdots$ is formally known as the **impulse-response function** of the VAR model.

There are several ways to estimate the impulse-response function. A simple, brute force method consists in using the companion form (4.12.28). Indeed, it follows directly from (4.12.29) that $\partial \gamma_{t+h} / \partial \zeta_t = F^h$. Since y_{t+h} and ε_t represent respectively the first n rows of γ_{t+h} and ζ_t, one can obtain Φ_h by simply retaining the first n rows and columns of F^h.

A more efficient way consists in obtaining the impulse-response function by numerical simulation. To do so, we may use the general VAR formulation (4.11.1), or better, its compact form equivalent (4.11.2). To simulate the impulse, implement recursively the following system for $h = 0, 1, 2, \dots$:

$$Y_h = X_h \mathcal{B} + \mathcal{E}_h \tag{4.13.2}$$

with:

$$Y_{-1} = \cdots = Y_{-p} = 0_{n\times n} \qquad X_h = \begin{pmatrix} 0_{n\times m} & Y_{h-1} & \cdots & Y_{h-p} \end{pmatrix} \qquad \mathcal{E}_0 = I_n \qquad \mathcal{E}_h = 0_{n\times n} \text{ for } h \neq 0 \tag{4.13.3}$$

[1] The lag polynomial introduced here is a multivariate generalization of the scalar lag polynomial defined in (3.9.56). Readers unfamiliar with the notion of inverse lag polynomial may safely skip this specific point and treat $A(L)^{-1}Cz_t$ as some constant.

The matrix $\mathcal{E}_0 = I_n$ implements the unit shocks a period t, and the system is then simulated for periods $t, t+1, t+2 \ldots$ The impulse-response function directly obtains from:

$$\Phi_h = Y_h' \tag{4.13.4}$$

In a Bayesian context, we must account for the fact that the impulse-response function $\Phi_0, \Phi_1, \Phi_2 \ldots$ are not constant parameters but random variables. Fortunately, it is easy to derive their posterior distributions by integrating them into a Gibbs sampler framework. Formally, we can derive a numerical approximation of the posterior distributions from the following algorithm:

algorithm 13.1: Gibbs sampling algorithm for the impulse-response function

1. at iteration j, draw $\beta^{(j)}$ from its posterior distributions. Recycle the values obtained from the j^{th} iteration of the Gibbs sampling algorithm. Reshape it into $\mathcal{B}^{(j)}$.
2. simulate the impulse-response function $\Phi_0^{(j)}, \Phi_1^{(j)}, \Phi_2^{(j)} \ldots$ from (4.13.2)-(4.13.4).
3. repeat until the desired number of iterations is realised.

13.2 Structural identification

Impulse-response functions aim at answering questions of the kind: "What is the effect of a unit shock in $\varepsilon_{i,t}$, everything else being held constant?". However, the shocks we are dealing with are typically correlated, i.e. the variance-covariance matrix of the residuals Σ is typically not diagonal (see equation (4.11.1)). As a consequence, the regular impulse-response functions may not constitute reliable estimates of the impact of a shock considered in isolation.

To solve this problem, we need to introduce the concept of structural VAR:

definition 13.1: a **structural VAR** is a model of the form:
$H_0 y_t = G z_t + H_1 y_{t-1} + \cdots + H_p y_{t-p} + \xi_t \qquad \xi_t \sim N(0, \Gamma) \qquad t = 1, \cdots, T$

There are two main differences between a structural VAR and the regular VAR (4.11.1) (the latter is also known as a **reduced-form VAR**). First, the structural VAR allows for contemporaneous interractions between the endogenous variables through the matrix H_0. Second, and most importantly, the variance-covariance matrix Γ is assumed to be diagonal so that the structural shocks ξ_t are uncorrelated. Following, the impulse-response functions produced by a structural VAR are meaningful, unlike those obtained from a reduced-form VAR.

It is straightforward to obtain a correspondance between a structural and a reduced-from VAR. Multiply both sides of the structural VAR in definition 13.1 by $H \equiv H_0^{-1}$ to obtain:

$$y_t = H\, G z_t + H\, H_1 y_{t-1} + \cdots + H\, H_p y_{t-p} + H\, \xi_t \tag{4.13.5}$$

Comparing with the reduced-form VAR (4.11.1), one directly obtains:

$$C = HG \qquad A_i = HH_i \qquad \varepsilon_t = H\xi_t \tag{4.13.6}$$

Also, another important relation obtains by noting that $\Sigma = \mathbb{E}(\varepsilon_t \varepsilon_t') = H\,\mathbb{E}(\xi_t \xi_t')H' = H\Gamma H'$, so that:

$$\Sigma = H\Gamma H' \tag{4.13.7}$$

The matrix H is known as a **structural identification matrix**. It permits the identification of the structural shocks ξ_t from the reduced-form shocks ε_t, and creates a correspondance between the reduced-form covariance matrix Σ and the structural shock covariance matrix Γ.

It is also straightforward to obtain the structural impulse response function from the structural identification matrix H. Rewrite the Wold representation in theorem 13.1 as:

$$y_t = A(L)^{-1}Cz_t + \Phi_0 HH^{-1}\varepsilon_t + \Phi_1 HH^{-1}\varepsilon_{t-1} + \Phi_2 HH^{-1}\varepsilon_{t-2}\dots \tag{4.13.8}$$

Or, using (4.13.6):

$$y_t = A(L)^{-1}Cz_t + \Psi_0\xi_t + \Psi_1\xi_{t-1} + \Psi_2\xi_{t-2}\dots \qquad \Psi_h \equiv \Phi_h H \qquad \xi_t \equiv H^{-1}\varepsilon_t \tag{4.13.9}$$

The series $\Psi_0, \Psi_1, \Psi_2\dots$ represents the **structural impulse-response function** of the model.

It should be clear that the structural identification matrix H constitutes the key element for structural identification. Once H is known, one can estimate a standard reduced form VAR, and then use H to recover the structural shocks ξ_t, the structural covariance matrix Γ, and the structural impulse-response function $\Psi_0, \Psi_1, \Psi_2\dots$. However, if a structural VAR uniquely defines the corresponding reduced-form VAR, the converse is not true. For a given reduced-form VAR, there exist in fact an infinite number of matrices H that satisfy (4.13.6) and (4.13.7), and thus an infinite number of possible structural VARs. Structural identification must then be conducted by reducing the set of possible candidates H with some economic theory.

Structural identification represents in fact a large field of research in economics, and finding a relevant matrix H can go far in sophistication. We cover some advanced approaches in chapter 14, but for now we limit the analysis to the introduction of two simple and classical identification strategies: **Cholesky factorization**, and **triangular factorization**.

With Cholesky factorization, the structural shocks are assumed to have unit variance, which implies that $\Gamma = I_n$. Following, (4.13.7) simplifies to:

$$\Sigma = HH' \tag{4.13.10}$$

Typically, we want to use an identification scheme that uniquely defines the structural matrix H. Because H is $n \times n$, it has n^2 elements to estimate and thus n^2 constraints on H are required to ensure a unique identification. (4.13.10) provides $n(n+1)/2$ restrictions, roughly half of the necessary restrictions. The remaining $n(n-1)/2$ restrictions are obtained by assuming that H is lower triangular, i.e. that the entries above its main diagonal are equal to 0.

To see what this implies in terms of economic interpretation, note that $\Psi_0 = H$ and so the structural Wold representation (4.13.9) writes $y_t = A(L)^{-1}Cz_t + H\xi_t + \Psi_1\xi_{t-1} + \Psi_2\xi_{t-2}\dots$ [2] Considering only the impact of ξ_t in the representation, we obtain:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{n,t} \end{pmatrix} = \dots + \begin{pmatrix} h_{11} & 0 & \cdots & 0 \\ h_{21} & h_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} \xi_{1,t} \\ \xi_{2,t} \\ \vdots \\ \xi_{n,t} \end{pmatrix} + \cdots \tag{4.13.11}$$

Therefore, $y_{1,t}$ is contemporaneously impacted by $\xi_{1,t}$, but not by the other structural shocks $\xi_{2,t}, \cdots, \xi_{n,t}$. $y_{2,t}$ is contemporaneously affected by $\xi_{1,t}$ and $\xi_{2,t}$, but not by $\xi_{3,t}, \cdots, \xi_{n,t}$, and so on. This kind of structural identification scheme thus implements **short-term restrictions** by stating that some of the structural shocks have no instantaneous effect on some of the model variables. It should also be clear from (4.13.11) that the ordering of the variables in the VAR becomes important since different

[2] Indeed, $\Phi_0 = I_n$ by construction. Following, (4.13.9) implies $\Psi_0 = \Phi_0 H = I_n H = H$.

restrictions apply to different variables. It is thus necessary to find a variable ordering that yields meaningful economic interpretation given the implied set of short-term restrictions.

With this setting, the structural identification exercise thus consists in finding a structural idenfication matrix H such that $\Sigma = HH'$, with H some lower triangular matrix. But this is precisely the definition of the Cholesky factor of Σ, and by property m.29, this factorization exists and is unique. Therefore, the structural VAR can be recovered directly from the reduced-form VAR, obtaining H trivially from Cholesky factorization of Σ.

In a Bayesian context, the structural impulse-response functions $\Psi_0, \Psi_1, \Psi_2 \ldots$ are also treated as random variables. It is straightforward to expand the Gibbs sampler methodology to structural VAR identification with Cholesky factorization, as summarized in the following algorithm:

algorithm 13.2: Gibbs sampling algorithm for structural identification, Cholesky factorization

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. simulate the impulse-response function $\Phi_0^{(j)}, \Phi_1^{(j)}, \Phi_2^{(j)} \ldots$ from (4.13.2)-(4.13.4).
3. obtain the structural matrix $H^{(j)}$ from the Cholesky factor of $\Sigma^{(j)}$.
4. calculate the structural impulse-response function $\Psi_0^{(j)}, \Psi_1^{(j)}, \Psi_2^{(j)} \ldots$ from $\Psi_h^{(j)} = \Phi_h^{(j)} H^{(j)}$.
5. repeat until the desired number of iterations is realised.

The assumption that the structural shocks have unit variance is sometimes too restrictive. In this case, one may instead carry structural identification by triangular factorization. With this scheme, the structural variance matrix Γ is diagonal but not necessarily identity. This creates n additional free parameters to estimate, and thus requires n additional constraints to ensure unicity of the identification. This is done by assuming that H is lower triangular, and additionally has its main diagonal made of ones. This creates short-term restrictions similar to the Cholesky structural identification, but imposes the further constraint of unit contemporaneous responses of the variables to their own structural shocks.

Structural identification thus amounts to finding a diagonal matrix Γ and a unit lower triangular matrix H such that $\Sigma = H\Gamma H'$. This is the definition of triangular factorization, and by property m.30 this factorization exists and is unique. Following, the Gibbs sampling algorithm can be directly adapted to the triangular factorization setting to yield:

algorithm 13.3: Gibbs sampling algorithm for structural identification, triangular factorization

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. simulate the impulse-response function $\Phi_0^{(j)}, \Phi_1^{(j)}, \Phi_2^{(j)} \ldots$ from (4.13.2)-(4.13.4).
3. obtain the variance matrix $\Gamma^{(j)}$ and the structural matrix $H^{(j)}$ from triangular factorization of $\Sigma^{(j)}$.
4. calculate the structural impulse-response function $\Psi_0^{(j)}, \Psi_1^{(j)}, \Psi_2^{(j)} \ldots$ from $\Psi_h^{(j)} = \Phi_h^{(j)} H^{(j)}$.
5. repeat until the desired number of iterations is realised.

Cholesky factorization and triangular factorizatin produce similar impulse-response function, up to a scaling term. With triangular factorization, ψ_{ij}^h provides the response to a unit shock (a structural shock $\xi_{j,t}$ of size 1), while with Cholesky factorization ψ_{ij}^h gives the response to a shock of one standard deviation (i.e. a structural shock $\xi_{j,t}$ of size $\sqrt{\gamma_j}$, where γ_j is the j^{th} diagonal term of Γ as obtained with triangular factorization).

13.3 Prediction

Prediction is a central concern in VAR modelling. It consists in predicting $\hat{y}_{T+1}, \hat{y}_{T+2}, \cdots, \hat{y}_{T+h}$ from a VAR model estimated with the sample $y_1, y_2, \cdots, y_T$. In a frequentist framework, computing forecasts involves the calculation of a minimum Mean Squared Error predictor. It can be shown (see e.g. Lütkepohl (2005), chapter 2) that the forecasts can be obtained recursively, taking conditional expectations of the general VAR formulation (4.11.1):

$$\begin{aligned}
\hat{y}_{T+1} &= C\hat{z}_{T+1} + A_1 y_T + \cdots + A_p y_{T-p+1} \\
\hat{y}_{T+2} &= C\hat{z}_{T+2} + A_1 \hat{y}_{T+1} + \cdots + A_p y_{T-p+2} \\
&\vdots
\end{aligned} \tag{4.13.12}$$

A confidence interval for $\hat{y}_{T+h}$ can then be obtained from:

$$\hat{y}_{T+h} \pm N_{\alpha/2}\, s_h \qquad s_h = \sqrt{diag(Q_h)} \qquad Q_h = Q_{h-1} + \Phi_h \hat{\Sigma} \Phi_h' \qquad Q_1 = \hat{\Sigma} \tag{4.13.13}$$

where Φ_h denotes the impulse response function matrix in (4.13.1).

In a Bayesian context, forecasts are formed using the posterior predictive distribution. As we want to predict $\hat{y}_{T+1}, \hat{y}_{T+2}, \cdots, \hat{y}_{T+h}$, the predictive distribution is given by:

$$f(\hat{y}_{T+1}, \cdots, \hat{y}_{T+h} | y) = \int f(\hat{y}_{T+1}, \cdots, \hat{y}_{T+h} | y, \theta)\ \pi(\theta | y)\ d\theta \tag{4.13.14}$$

Unlike the linear regression model however, analytical expressions are generally unavailable for the predictive distribution when predictions are considered beyond one period. Therefore, one must compute (4.13.14) by the way of simulation methods, which is easily done from algorithm 6.3. Noting that $\theta = \{\beta, \Sigma\}$ for Bayesian VAR models, it is straightforward to adapt the algorithm that becomes:

algorithm 13.4: Gibbs sampling algorithm for the posterior predictive distribution

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. Recover $C^{(j)}, A_1^{(j)}, \cdots, A_p^{(j)}$ from $\beta^{(j)}$.
2. draw $\varepsilon_{T+1}, \cdots, \varepsilon_{T+h}$ from $\varepsilon_t \sim N(0, \Sigma)$.
3. generate recursively $\hat{y}_{T+1}, \cdots, \hat{y}_{T+h}$ from:
$$\begin{aligned}
\hat{y}_{T+1}^{(j)} &= C\hat{z}_{T+1} + A_1 y_T + \cdots + A_p y_{T+1-p} + \varepsilon_{T+1} \\
\hat{y}_{T+2}^{(j)} &= C\hat{z}_{T+2} + A_1 \hat{y}_{T+1} + \cdots + A_p y_{T+2-p} + \varepsilon_{T+2} \\
&\vdots \\
\hat{y}_{T+h}^{(j)} &= C\hat{z}_{T+h} + A_1 \hat{y}_{T+h-1} + \cdots + A_p \hat{y}_{T+h-p} + \varepsilon_{T+h}
\end{aligned}$$
4. marginalize, that is, discard $\beta^{(j)}$ and $\Sigma^{(j)}$ and keep only the predictions $\hat{y}_{T+1}^{(j)}, \cdots, \hat{y}_{T+h}^{(j)}$.
5. repeat until the desired number of iterations is realised.

Two points are worth noting with algorithm 13.4. First, the algorithm builds recursively, and thus involves both the observed values $y_T, y_{T-1}, \cdots$ and the predicted values $\hat{y}_{T+1}, \hat{y}_{T+2}, \ldots$ in the construction of the forecasts. Second, the forecasts involve the predicted values for the exogenous $\hat{z}_{T+1}, \cdots, \hat{z}_{T+h}$. As these predicted values are not obtained from the model, they need to be exogenously supplied by the researcher.

Once predictions are obtained, forecast evaluation criteria can be calculated to assess the predictive performance of the model. Most formulas presented here are direct adaptations of the regression formulas,

adapted to VAR modelling. We start with classic in-sample quantities. Denote by $\hat{\mathcal{B}}$ the posterior median of $\mathcal{B}$, and by $\hat{\mathcal{E}}$ the resulting median residual obtained from $\hat{\mathcal{E}} = Y - X\hat{\mathcal{B}}$, using the compact VAR formulation (4.11.2). The following quantities obtain:

$$SSR = diag(\hat{\mathcal{E}}'\hat{\mathcal{E}}) \qquad TSS = diag((Y-\bar{Y})'(Y-\bar{Y})) \qquad R^2 = 1 - \frac{SSR}{TSS} \qquad \text{adj-R}^2 = 1 - (1-R^2)\frac{T-1}{T-k} \tag{4.13.15}$$

where the results are n-dimensional vectors, and operations are conducted elementwise. For the maximum likelihood VAR model, additional in-sample lag criteria are provided by the Akaike Information Criterion (AIC), the Schwarz's Bayesian Information Criterion (BIC), and the Hannan-Quinn criterion (HQ), respectively defined as:

$$AIC = 2q/T - 2\,\hat{L}/T \qquad BIC = q\log(T)/T - 2\,\hat{L}/T \qquad HQ = 2q\,\log(\log(T))/T - 2\,\hat{L}/T \tag{4.13.16}$$

with $\hat{L}$ 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. 75), these criteria rewrite as:

$$AIC = 2q/T + \log(|\hat{\Sigma}|) \qquad BIC = q\,\log(T)/T + \log(|\hat{\Sigma}|) \qquad HQ = 2q\,\log(\log(T))/T + \log(|\hat{\Sigma}|) \tag{4.13.17}$$

Standard out-of-sample forecast evaluation criteria can be readily adapted to VAR models, such as the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). For a forecast up to y_{T+h}, they are defined as:

$$RMSE = \sqrt{\frac{1}{h}\sum_{j=1}^{h}(y_{i,T+j} - \hat{y}_{i,T+j})^2} \qquad MAE = \frac{1}{h}\sum_{j=1}^{h}|y_{i,T+j} - \hat{y}_{i,T+j}| \qquad MAPE = \frac{100}{h}\sum_{j=1}^{h}\left|\frac{y_{i,T+j} - \hat{y}_{i,T+j}}{y_{i,T+j}}\right| \tag{4.13.18}$$

where $y_{i,T+j}$ and $\hat{y}_{i,T+j}$ respectively denote the actual and predicted values for variable i at period $T+j$. Similarly, The Theil inequality coefficient (Theil-U) and bias are given by:

$$\text{Theil-U} = \frac{\sqrt{\sum_{j=1}^{h}(y_{i,T+j} - \hat{y}_{i,T+j})^2}}{\sqrt{\sum_{j=1}^{h} y_{i,T+j}^2} + \sqrt{\sum_{j=1}^{h} \hat{y}_{i,T+j}^2}} \qquad \text{bias} = \frac{\sum_{j=1}^{h} y_{i,T+j} - \hat{y}_{i,T+j}}{\sum_{j=1}^{h}|y_{i,T+j} - \hat{y}_{i,T+j}|} \tag{4.13.19}$$

The log score (LogS) and the continuous ranked probability score (CRPS) are defined as:

$$LogS = -\,\log(\hat{f}(y_{i,T+j})) \qquad CRPS = \int_{-\infty}^{+\infty}\left[\hat{F}(z) - \mathbb{1}(y_{i,T+j} \leq z)\right]^2 dz \tag{4.13.20}$$

For the log score, we follow the strategy suggested by Warne et al. (2013) and use 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:

$$LogS = -\,\log(\hat{\phi}(y_{i,T+j})) \tag{4.13.21}$$

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 variable i at period $T+j$. To evaluate the log score for all prediction periods jointly, we treat $\hat{y}_{i,T+1}, \cdots, \hat{y}_{i,T+h}$ as a joint multivariate normal distribution and evaluate their empirical mean and variance-covariance matrix from the Gibbs sampler draws, then evaluate the log score from the multivariate normal density.

For the CRPS, we follow again Krüger et al. (2017) and use the consistent approximation:

$$CRPS \approx \frac{1}{J}\sum_{j=1}^{J}|\hat{y}_{i,t+h}^{(j)} - y_{i,t+h}| - \frac{1}{2\,J^2}\sum_{j=1}^{J}\sum_{k=1}^{J}|\hat{y}_{i,t+h}^{(j)} - \hat{y}_{i,t+h}^{(k)}| \tag{4.13.22}$$

A joint CRPS over all prediction periods can be obtained as the sum of the individual CRPS for $\hat{y}_{i,T+1}, \cdots, \hat{y}_{i,T+h}$.

13.4 Forecast error variance decomposition

Another useful application related to prediction is the so-called forecast error variance decomposition. It determines the contribution of each shock of the model to the forecast error, the unpredictable component of the forecast. It thus explains which shocks matter to explain a variable at different forecast horizons.

Assume we want to generate a prediction for period $t+h$. To do so we start from the structural Wold representation (4.13.9), evaluated at $t+h$.

$$y_{t+h} = A(L)^{-1}Cz_{t+h} + \Psi_0\xi_{t+h} + \Psi_1\xi_{t+h-1} + \Psi_2\xi_{t+h-2}\dots \tag{4.13.23}$$

Considering a prediction made at period t, this expression can be separated into three components:

$$y_{t+h} = A(L)^{-1}Cz_{t+h} + \sum_{i=0}^{\infty}\Psi_{h+i}\xi_{t-i} + \sum_{i=0}^{h-1}\Psi_i\xi_{t+h-i} \tag{4.13.24}$$

On the right-hand side, the first term represents the exogenous component at $T+h$ (assumed to be known for prediction pruposes), while the second term contains the known present and past shocks. The final term contains the future shocks. Following, the forecast for period $t+h$ is given by:

$$\mathbb{E}(y_{t+h}) = A(L)^{-1}Cz_{t+h} + \sum_{i=0}^{\infty}\Psi_{h+i}\xi_{t-i} \tag{4.13.25}$$

And thus the forecast error is given by:

$$y_{t+h} - \mathbb{E}(y_{t+h}) = \sum_{i=0}^{h-1}\Psi_i\xi_{t+h-i} \tag{4.13.26}$$

Developing:

$$\begin{pmatrix} y_{1,t+h} - \mathbb{E}(y_{1,t+h}) \\ y_{2,t+h} - \mathbb{E}(y_{2,t+h}) \\ \vdots \\ y_{n,t+h} - \mathbb{E}(y_{n,t+h}) \end{pmatrix} = \begin{pmatrix} \psi_{11}^0 & \psi_{12}^0 & \cdots & \psi_{1n}^0 \\ \psi_{21}^0 & \psi_{22}^0 & \cdots & \psi_{2n}^0 \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{n1}^0 & \psi_{n2}^0 & \cdots & \psi_{nn}^0 \end{pmatrix} \begin{pmatrix} \xi_{1,t+h} \\ \xi_{2,t+h} \\ \vdots \\ \xi_{n,t+h} \end{pmatrix} + \begin{pmatrix} \psi_{11}^1 & \psi_{12}^1 & \cdots & \psi_{1n}^1 \\ \psi_{21}^1 & \psi_{22}^1 & \cdots & \psi_{2n}^1 \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{n1}^1 & \psi_{n2}^1 & \cdots & \psi_{nn}^1 \end{pmatrix} \begin{pmatrix} \xi_{1,t+h-1} \\ \xi_{2,t+h-1} \\ \vdots \\ \xi_{n,t+h-1} \end{pmatrix} + \cdots \tag{4.13.27}$$

So the forecast error for variable i rewrites as:

$$y_{i,t+h} - \mathbb{E}(y_{i,t+h}) = \sum_{j=0}^{h-1}\psi_{i1}^j\xi_{1,t+h-j} + \sum_{j=0}^{h-1}\psi_{i2}^j\xi_{2,t+h-j} + \cdots + \sum_{j=0}^{h-1}\psi_{in}^j\xi_{n,t+h-j} \tag{4.13.28}$$

Denote the variance of this forecast error by $\sigma_{y_i}(h)$, and denote by $\gamma_1, \cdots, \gamma_n$ the variances of the structural shocks found on the diagonal of Γ. Then taking variances on both sides of (4.13.28) and noting that no

covariances are implid since the structural shocks are uncorrelated, one obtains:

$$\sigma_{y_i}(h) = \gamma_1 \sum_{j=0}^{h-1} (\psi_{i1}^j)^2 + \gamma_2 \sum_{j=0}^{h-1} (\psi_{i2}^j)^2 + \cdots + \gamma_n \sum_{j=0}^{h-1} (\psi_{in}^j)^2 \tag{4.13.29}$$

Eventually, divide both sides by $\sigma_{y_i}(h)$ to obtain:

$$1 = \underbrace{\frac{\gamma_1}{\sigma_{y_i}(h)} \sum_{j=0}^{h-1} (\psi_{i1}^j)^2}_{\substack{\text{Contribution of shock 1} \\ \text{in forecast error}}} + \underbrace{\frac{\gamma_2}{\sigma_{y_i}(h)} \sum_{j=0}^{h-1} (\psi_{i2}^j)^2}_{\substack{\text{Contribution of shock 2} \\ \text{in forecast error}}} + \cdots + \underbrace{\frac{\gamma_n}{\sigma_{y_i}(h)} \sum_{j=0}^{h-1} (\psi_{in}^j)^2}_{\substack{\text{Contribution of shock n} \\ \text{in forecast error}}} \tag{4.13.30}$$

Or:

$$1 = \sigma_{y_i}^1(h) + \sigma_{y_i}^2(h) + \cdots + \sigma_{y_i}^n(h) \qquad \sigma_{y_i}^k(h) \equiv \frac{\gamma_k}{\sigma_{y_i}(h)} \sum_{j=0}^{h-1} (\psi_{ik}^j)^2 \qquad k = 1, \cdots, n \tag{4.13.31}$$

$\sigma_{y_i}^1(h), \cdots, \sigma_{y_i}^n(h)$ constitute the **forecast error variance decomposition** of the model. $\sigma_{y_i}^k(h)$ represents the proportion of forecast error variance of variable i due to structural shock k at horizon $t+h$. It thus indicates the importance of shock k to explain (or predict) variable i at the forecast horizon $t+h$. Typically, the forecast error variance of a variable is explained by its own shocks at short horizons and by shocks to other variables at longer horizons.

In a Bayesian context the parameters $\sigma_{y_i}^1(h), \cdots, \sigma_{y_i}^n(h)$ are treated as random variables. Since the posteriors don't have analytical solutions, estimation must be integrated to Gibbs sampler framework. This yields the following algorithm:

algorithm 13.5: Gibbs sampling algorithm for forecast error variance decomposition

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. simulate the impulse-response function $\Phi_0^{(j)}, \Phi_1^{(j)}, \Phi_2^{(j)} \ldots$ from (4.13.2)-(4.13.4).
3. obtain the structural matrix $H^{(j)}$ and the structural variance matrix $\Gamma^{(j)}$ from $\Sigma^{(j)}$.
4. calculate the structural impulse-response function $\Psi_0^{(j)}, \Psi_1^{(j)}, \Psi_2^{(j)} \ldots$ from $\Psi_h^{(j)} = \Phi_h^{(j)} H^{(j)}$.
5. calculate $\sigma_{y_i}^1(h)^{(j)}, \cdots, \sigma_{y_i}^n(h)^{(j)}$ from $\sigma_{y_i}^k(h)^{(j)} = \frac{\gamma_k}{\sigma_{y_i}(h)} \sum_{j=0}^{h-1} (\psi_{ik}^j)^2$
6. repeat until the desired number of iterations is realised.

13.5 Historical decomposition

Forecast error variance decomposition is concerned with the contribution of structural shocks for predictions. It may also be interesting to establish the contribution of the structural shocks in the past, over the observed data sample. This is the purpose of historical decomposition.

Consider again the structural Wold representation (4.13.9), considered at any sample period t:

$$y_t = A(L)^{-1} C z_t + \Psi_0 \xi_t + \Psi_1 \xi_{t-1} + \Psi_2 \xi_{t-2} + \cdots \qquad t = 1, \cdots, T \tag{4.13.32}$$

The right-hand side can be decomposed into two components. The first term represents the deterministic part of the model, while the remaining terms give the contribution of the unpredictable structural

disturbances affecting the dynamics of the model. Using the notation $d_t \equiv A(L)^{-1}Cz_t$ to designate the deterministic part of the model, the developed representation writes:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{n,t} \end{pmatrix} = \begin{pmatrix} d_{1,t} \\ d_{2,t} \\ \vdots \\ d_{n,t} \end{pmatrix} + \begin{pmatrix} \psi_{11}^0 & \psi_{12}^0 & \cdots & \psi_{1n}^0 \\ \psi_{21}^0 & \psi_{22}^0 & \cdots & \psi_{2n}^0 \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{n1}^0 & \psi_{n2}^0 & \cdots & \psi_{nn}^0 \end{pmatrix} \begin{pmatrix} \xi_{1,t} \\ \xi_{2,t} \\ \vdots \\ \xi_{n,t} \end{pmatrix} + \begin{pmatrix} \psi_{11}^1 & \psi_{12}^1 & \cdots & \psi_{1n}^1 \\ \psi_{21}^1 & \psi_{22}^1 & \cdots & \psi_{2n}^1 \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{n1}^1 & \psi_{n2}^1 & \cdots & \psi_{nn}^1 \end{pmatrix} \begin{pmatrix} \xi_{1,t-1} \\ \xi_{2,t-1} \\ \vdots \\ \xi_{n,t-1} \end{pmatrix} + \cdots \tag{4.13.33}$$

Following, we can rewrite the representation by grouping the shocks as:

$$y_{i,t} = d_{i,t} + \underbrace{\sum_{j=0}^{\infty} \psi_{i1}^j \xi_{1,t-j}}_{\substack{\text{Historical contribution} \\ \text{of shock 1}}} + \underbrace{\sum_{j=0}^{\infty} \psi_{i2}^j \xi_{2,t-j}}_{\substack{\text{Historical contribution} \\ \text{of shock 2}}} + \cdots + \underbrace{\sum_{j=0}^{\infty} \psi_{in}^j \xi_{n,t-j}}_{\substack{\text{Historical contribution} \\ \text{of shock n}}} \tag{4.13.34}$$

Or:

$$y_{i,t} = d_{i,t} + h_{i1,t} + h_{i2,t} + \cdots + h_{in,t} \qquad h_{ik,t} \equiv \sum_{j=0}^{\infty} \psi_{ik}^j \xi_{k,t-j} \tag{4.13.35}$$

Representation (4.13.35) provides the **historical decomposition** of the model. Each $h_{ik,t}$ on the right-hand side represents the historical contributions of structural shock k of the model in the historical value of variable i. Note that the summations are of infinite order. In practice however the sample used for estimation is of finite size and comprise only T observations. Therefore for any $t = 1, \cdots, T$, the historical decomposition actually obtains from:

$$h_{ik,t} = \sum_{j=0}^{t-1} \psi_{ik}^j \xi_{k,t-j} \tag{4.13.36}$$

(4.13.36) implies that for small t, only few terms of the infinite order structural Wold representation are effectively involved in the calculation of the historical decomposition. This implies that historical decomposition is typically more accurately estimated at the end of the sample, while only rough approximations are obtained for the first sample periods.

In a Bayesian context the structural Wold representation parameters are random variables, so historical decomposition must be as usual integrated to Gibbs sampler framework. This yields the following algorithm:

algorithm 13.6: Gibbs sampling algorithm for historical decomposition

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. simulate the impulse-response function $\Phi_0^{(j)}, \Phi_1^{(j)}, \Phi_2^{(j)} \ldots$ from (4.13.2)-(4.13.4).
3. obtain the structural matrix $H^{(j)}$ from $\Sigma^{(j)}$.
4. calculate the structural impulse-response function $\Psi_0^{(j)}, \Psi_1^{(j)}, \Psi_2^{(j)} \ldots$ from $\Psi_h^{(j)} = \Phi_h^{(j)} H$.
5. obtain the structural shocks $\xi_t^{(j)}, \xi_{t-1}^{(j)}, \xi_{t-2}^{(j)} \ldots$ from $\xi_t^{(j)} = H^{-1}\varepsilon_t^{(j)}$.
6. calculate $h_{i1,t}^{(j)}, \cdots, h_{in,t}^{(j)}$ from $h_{ik,t}^{(j)} = \sum_{j=0}^{t-1} \psi_{ik}^j \xi_{k,t-j}$.
7. calculate $d_{i,t}^{(j)} = y_{i,t} - h_{i1,t} - h_{i2,t} - \cdots - h_{in,t}$.
8. repeat until the desired number of iterations is realised.

13.6 Application: how well does the IS-LM model fit postwar E.U. data?

In a seminal paper, Gali (1992) investigates whether postwar U.S. data supports the stylized predictions of the canonical IS-LM model. To do so the study estimates a simple, four-variable VAR model, then confronts its empirical results to the following stylized IS-LM predictions:

1. Positive supply shocks, real demand shocks and money supply shocks have (at least) a transitory positive effect on GDP growth.

2. Positive supply and real demand shocks both result in higher GDP growth, but have opposite effects on inflation.

3. Short-term economic fluctuations are mostly driven by real demand and monetary shocks, while supply shocks take over in the long run.

4. The short-term interest rate declines after a positive money supply shock, but increases following a positive money demand shock.

5. Monetary shocks are transmitted to the real sector through their impacts on the interest rate.

The paper finds that the estimated VAR overall agrees with these stylized facts, up to a few oddities such as the prevalence of supply shocks in the variation of output at short horizon.

This section replicates the same exercise, using this time Euro Area data instead of U.S. data. The VAR model includes the same four variables as in Gali (1992): real GDP growth, broad money m3 as yearly growth rate, the 3-month interest rate, and CPI inflation. The data comes from Eurostats, the OECD and the European Central Bank. The sample is quarterly and covers years 1974 to 2024, which is substantially longer than its U.S. counterpart. Whenever data is missing for early years, it is supplemented by the excellent Area-Wide Model dataset of Fagan et al. (2001). The dataset is represented in Figure 13.1:

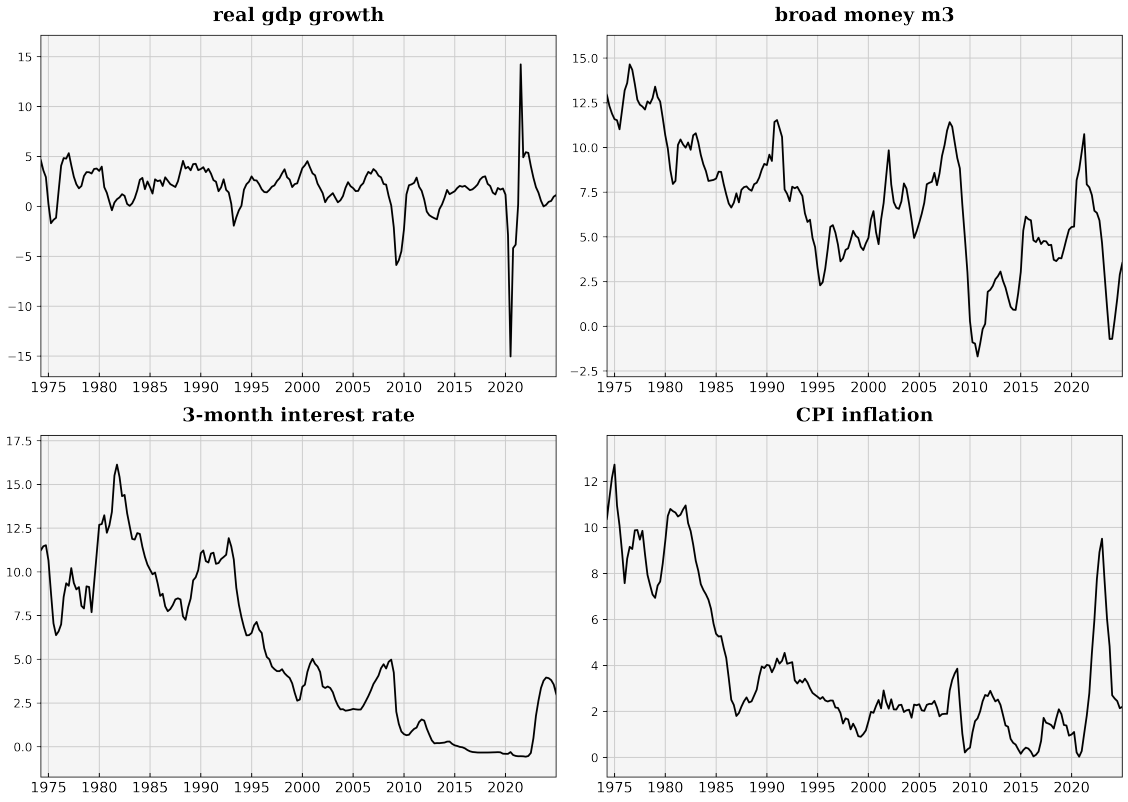


Figure 13.1: IS-LM dataset for the Euro Area

This section focuses on two aspects. First, it estimates a simple Bayesian VAR model similar to that of Gali (1992), and tries to establish whether the IS-LM stylized facts are also validated by Euro Area data. It does so by the way of a simple structural identification scheme and the analysis of impulse response function, forecast error variance decomposition and historical decomposition. Second, using the framework as a pretext, it tries to establish whether Bayesian VAR models perform better than their maximum likelihood counterpart in terms of predictive performance. To do so an expanding window exercise is conducted, with forecast evaluation at short and medium horizons.

For the first part of the exercise, a simple Bayesian model with four lags is estimated on the dataset. The retained prior is the normal-Wishart, but the results are fairly invariant to the selected prior. Similar to Gali (1992), the model aims at identifying four structural shocks assumed to drive the data dynamics: supply shocks, real demand shocks (also labelled "IS shocks"), and monetary shocks further divided into money supply and money demand shocks. Structural identification is conducted with the simplest possible setup: Cholesky factorisation. This amounts to setting short-term restrictions on the structural impulse response function of the model. Indeed, from (4.13.11), one obtains:

$$\begin{pmatrix} gdp_t \\ m3_t \\ rate_t \\ cpi_t \end{pmatrix} = \begin{pmatrix} h_{11} & 0 & 0 & 0 \\ h_{21} & h_{22} & 0 & 0 \\ h_{31} & h_{32} & h_{33} & 0 \\ h_{41} & h_{42} & h_{43} & h_{44} \end{pmatrix} \begin{pmatrix} \xi_t^s \\ \xi_t^{ms} \\ \xi_t^{md} \\ \xi_t^d \end{pmatrix} + \cdots \tag{4.13.37}$$

where ξ_t^s, ξ_t^{ms}, ξ_t^{md} and ξ_t^d respectively denote supply, money supply, money demand and real demand shocks. Ordering matters with Cholesky factorization, and the retained order implies the following assumptions:

1. the real sector (GDP growth) is only affected immediately by supply shocks. Shocks on the demand side only impact production with a lag due to the time required by producers to adapt to new demand conditions. This is a fairly usual assumption.

2. money growth responds instantaneously to supply and money supply shocks, but monetary authorities adjust with a lag to money demand and real demand conditions.

3. The short-term interest rate adjusts immediately to supply and monetary shocks, but takes one quarter to adjust to changes in real demand.

4. CPI inflation adjusts immediately to all disturbances, reflecting the fact that prices adjust continuously as a result to moves in supply and demand in all markets.

Figure 13.2 reports the structural impulse response obtained with this identification scheme. Overall, the responses agree with the IS-LM stylized facts. Positive supply, money supply and money demand shocks all have an expansionary effect on real GDP growth. The impact is by far the strongest for supply shocks, at 1.5 percent point at impact. The negative response to real demand shocks on the other hand is counter-intuitive and clearly at odd with IS-LM theory, though it is small and initially not significant.

As expected, broad money m3 increases following any shock on the demand side, at least in the medium run. The initial decline following a positive supply shock reflects the adjustment of money supply to lower nominal transactions, before increasing again due to higher economic activity.

CPI inflation responds postiviely to monetary shocks and real demand shocks, as expected. However, the positive response to supply shocks is probably the strongest contradiction with the IS-LM stylized facts. Instead of showing a drop in price level due to lower input prices, it seems to suggest a fueling in price, possibly due to the surge in economic activity resulting from the shock. This is anyway clearly at odd with standard IS-LM theory.

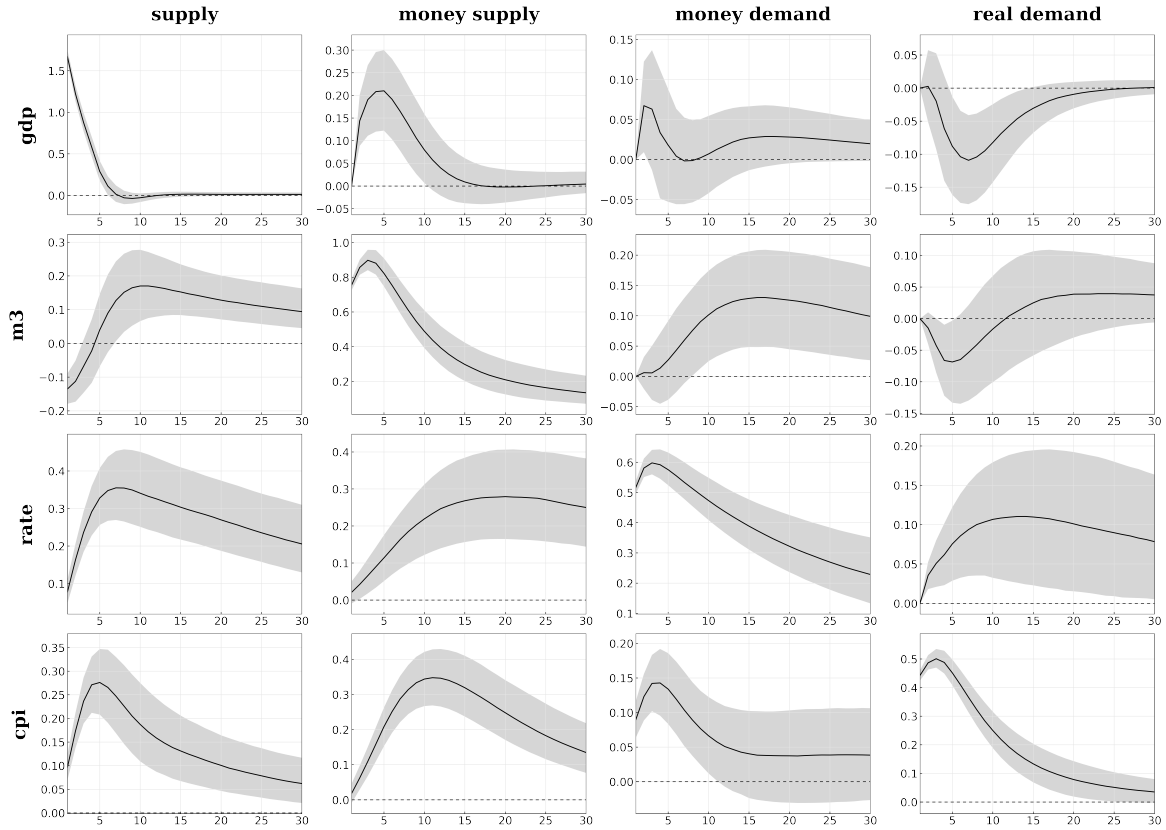


Figure 13.2: Structural impulse response function

As a result, in a Taylor rule fashion, the short-term interest rate adjusts upwards following a positive supply shock, a logical but nevertheless unexpected contradiction with the IS-LM framework. The upward response of the short-term rate to the other shocks on the other hand is consistent with traditional Keynesian views.

As a next step, the exercise considers the forecast error variance decomposition of the four variables included in the model. Figure 13.3 summarizes the estimates for the model. Unlike standard IS-LM wisdom, fluctuations in GDP can be seen to be dominated by supply shocks, even at business cycle horizons. In fact, supply shocks almost exclusively explain output fluctuations at any horizon. This is a critical contradiction with traditional Keynesian views that grant a significant role to demand shocks for stabilization purposes.

Broad money growth appears almost exclusively determined by money supply shocks, a view more aligned with monetarism and the quantitative theory of money than with traditional Keynesian beliefs that would emphasize the contribution of money demand and real demand shocks.

The short-term interest rate proves overall consistent with IS-LM predictions. At short horizon it is dominated by money demand shocks, before it leaves some space for supply and money supply shocks at longer horizons.

Finally, CPI inflation appears mostly determined by the real demand side of the economy at business cycle horizons, with money supply playing a larger role at long horizons. This facts are consistent with traditional Keynesian views on price adjustment.

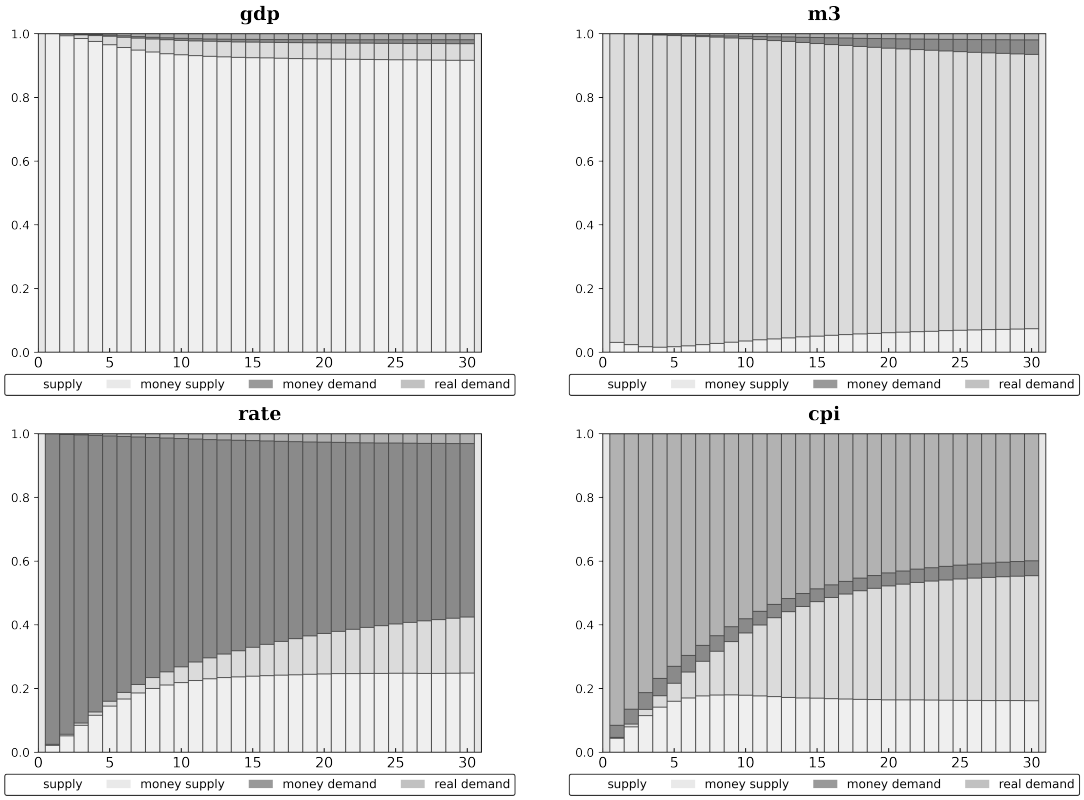


Figure 13.3: Forecast error variance decomposition

The final step in this first part examines the historical decomposition of the series, as displayed in Figure 13.4. The results overall support the preceding conclusions. The fluctuations of real GDP growth can be seen to be overhelmingly dominated by the contributions of supply shocks. This is especially true for the 2007 financial crisis and the 2020 pandemic crisis. At the margin, money supply shocks seem to contribute mostly negatively to output fluctuations. Perhaps paradoxically, the negative contributions often appear after recessions, like in 1994, 2010 and 2022, when the economy would yet require more accomodating monetary policies. Often though those negative contributions can be related to preceding inflationary pressures calling for monetary adjustments.

Unsurprisingly, broad money fluctuations are dominated by money supply components. Supply and money demand components sometimes play a small role, such as during the 2010 decade when the Euro area reached a liquidity trap.

Interestingly enough, the fluctuations in the short term interest rate display a mix of contributions across the period. Money supply and real demand shocks seem to dominate the pre-2000 sample, while subsequent years reveal a stronger contribution of supply shocks in the determination of monetary policy.

CPI inflation finally seem mostly determined by money supply and real demand shocks. Contributions from money demand are almost absent, here again a paradox regarding Keynesian views that grant a substantial role to money demand on the money market, and subsequently on the price level. Also, supply shock contributions seem fairly small in general. This is particularly interesting in the light of the recent surge in inflation following the 2020 pandemic. The rise in raw material and energy costs made some analysts conclude to a major contribution of supply shocks over the period, like Bernanke and Blanchard (2023). The model suggests on the contrary that real demand shocks played a central role in the episode, with further fueling emanating from positive money supply shocks. This alternative narrative is supported by other studies such as Giannone and Primiceri (2023) who advocate the role of expansionary fiscal policies, strong consumer demand following the pandemic, and accomodative monetary policies.

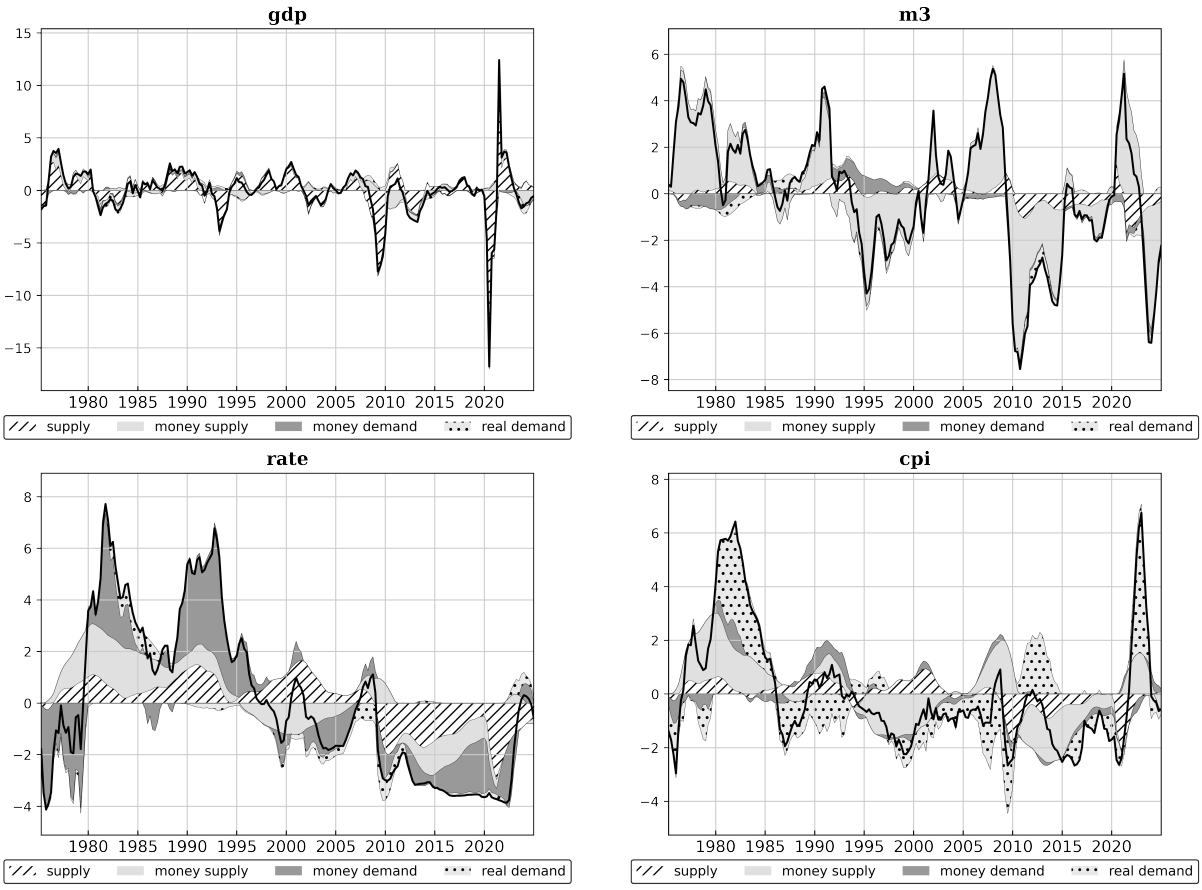


Figure 13.4: Historical decomposition

The second part of the exercise focuses on predictive performance. It aims at demonstrating the main point of Bayesian models, namely that they produce better forecasts than their maximum likelihood counterparts. To do so, an expanding window setup is developed. The initial sample covers the periods 1974Q1-2004Q1 and is then expanded one quarter at a time until 2023Q4, resulting in 80 expanding windows. For each window out-of-sample forecasts are conducted up to four periods ahead for all available VAR models, and forecast evluation criteria are calculated. The two retained criteria are the standard root mean squared error, and the Bayesian-specific log score. Table 13.1 reports the results for the two criteria at prediction horizon $t+1$ and $t+4$, averaged over all windows.

Looking first at the root mean squared errors, the maximum likelihood VAR consistently appears as the lowest performer. Typical difference with the Bayesian models is about 15%, with smaller spreads for CPI inflation and significantly higher differences for the short-term rate where the gap is almost 50% at horizon $t+1$. The conclusion is robust to the forecast horizon, implying Bayesian VAR models perform better both at short and medium terms. The predictive performance across Bayesian models then seems to favour the Minnesota, independent and large BVAR priors (the latter two being very similar) against the normal-Wishart and dummy observation priors. Overall yet the performance are quite close one from each other.

The main shortcoming of the RMSE is that it only considers point estimates while in real-life applications forecasting uncertainty also matters. The log score takes into account the full predictive density and thus provides a richer picture of the predictive performance. Looking at this criterion, the Minnesota becomes the weakest model, with significantly higher values for all variables but the short rate. This suggests that the Minnesota provides good point estimates, but poor prediction intervals. The large BVAR prior seems to dominate the other priors at the margin, but overall the last four Bayesian priors produce very similar predictive performance.

	gdp	m3	rate	cpi
maximum likelihood	1.410	0.696	0.316	0.452
Minnesota	1.135	0.642	0.211	0.404
normal-Wishart	1.166	0.658	0.236	0.410
independent	1.132	0.647	0.206	0.404
dummy observations	1.161	0.653	0.238	0.411
large bvar	1.138	0.649	0.205	0.409

(a) RMSE at $t+1$

	gdp	m3	rate	cpi
maximum likelihood	2.354	1.747	0.763	1.099
Minnesota	2.003	1.594	0.589	1.029
normal-Wishart	2.064	1.613	0.637	1.051
independent	1.978	1.596	0.572	1.025
dummy observations	2.065	1.612	0.637	1.049
large bvar	2.016	1.599	0.586	1.025

(b) RMSE at $t+4$

	gdp	m3	rate	cpi
maximum likelihood	-	-	-	-
Minnesota	4.643	1.375	0.441	1.064
normal-Wishart	3.978	1.332	0.571	0.944
independent	4.020	1.331	0.544	0.947
dummy observations	3.909	1.323	0.571	0.952
large bvar	3.976	1.323	0.571	0.905

(c) Log score at $t+1$

	gdp	m3	rate	cpi
maximum likelihood	-	-	-	-
Minnesota	4.761	3.009	1.575	3.129
normal-Wishart	4.053	2.827	1.664	2.601
independent	4.290	2.836	1.584	2.730
dummy observations	4.112	2.825	1.664	2.596
large bvar	4.155	2.809	1.622	2.521

(d) Log score at $t+4$

Table 13.1: Out-of-sample predictive performance

CHAPTER 14

Bayesian VAR: advanced applications

14.1 Conditional forecasts: an agnostic approach

This section introduces the notion **conditional forecasts**, which is closely related to the idea of scenario analysis. The basic approach developed in this section is very simple and follows Banbura et al. (2015). It consists in treating the forecasts as unobserved state variables and the conditions as observed variables, then integrate the whole setup in a standard Bayesian state-space framework.

Formally, assume that we have a VAR model and want to use it to generate forecasts $\hat{y}_t$ for periods $t = T+1, \cdots, T+h$. We further want to implement conditions, that is, we want to constrain the path of certain variables at certain forecast periods to take specific values exogenously decided. This means that for some variable(s) i ($i = 1, \cdots, n$) and some forecast period(s) t ($t = T+1, \cdots, T+h$), we want to set $\hat{y}_{i,t} = \bar{y}_{i,t}$, with $\bar{y}_{i,t}$ some value set exogenously for the scenario. The set of values $\bar{y}_{i,t}$ then represent the conditions for the exercise.

We may want the conditions to hold exactly, in which case they are called **hard conditions**. We may also want to allow for some variability around the conditions, in which case they are called **soft conditions**. A convenient way to represent this consists in assuming that the conditions are normally distributed random variables:

$$\hat{y}_{i,t} \sim N(\bar{y}_{i,t}, \omega_{i,t}) \qquad \Rightarrow \qquad \hat{y}_{i,t} = \bar{y}_{i,t} + \epsilon_{i,t} \qquad \epsilon_{i,t} \sim N(0, \omega_{i,t}) \tag{4.14.1}$$

We may represent hard conditions by setting $\omega_{ij} = 0$ and soft conditions with positive values for ω_{ij}. For the incoming developments, it is also useful to notice that we can represent the absence of conditions on $\hat{y}_{i,t}$ by setting $\bar{y}_{i,t} = 0$ and $\omega_{i,t}$ to a very large value, which amounts to setting a diffuse prior belief on $\hat{y}_{i,t}$. Gathering hard conditions, soft conditions and no-conditions for forecast period t in a single n-dimensional vector $\bar{y}_t$, one obtains:

$$\hat{y}_t = \bar{y}_t + \epsilon_t \qquad \epsilon_t \sim N(0, \Omega_t) \qquad \Omega_t = diag(\omega_t) \tag{4.14.2}$$

with:

$$\hat{y}_t = \begin{pmatrix} \hat{y}_{1,t} \\ \hat{y}_{2,t} \\ \vdots \\ \hat{y}_{n,t} \end{pmatrix} \qquad \bar{y}_t = \begin{pmatrix} \bar{y}_{1,t} \\ \bar{y}_{2,t} \\ \vdots \\ \bar{y}_{n,t} \end{pmatrix} \qquad \epsilon_t = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \vdots \\ \epsilon_{n,t} \end{pmatrix} \qquad \omega_t = \begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \\ \vdots \\ \omega_{n,t} \end{pmatrix} \tag{4.14.3}$$

Using the symmetry of the normal distribution, (4.14.2) rewrites:

$$\bar{y}_t = \hat{y}_t + \epsilon_t \qquad \epsilon_t \sim N(0, \Omega_t) \qquad \Omega_t = diag(\omega_t) \tag{4.14.4}$$

This represents the observation part of the setup. For the dynamic part, note that the forecasts obtain from the VAR model:

$$\hat{y}_t = Cz_t + A_1 \hat{y}_{t-1} + \cdots + A_p \hat{y}_{t-p} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \Sigma) \qquad t = T+1, \cdots, T+h \tag{4.14.5}$$

We can rewrite the model in companion form as in (4.12.27):

$$\hat{\gamma}_t = \mu_t + F\hat{\gamma}_{t-1} + \zeta_t \qquad \zeta_t \sim N(0,K) \tag{4.14.6}$$

with:

$$\hat{\gamma}_t = \begin{pmatrix} \hat{y}_t \\ \hat{y}_{t-1} \\ \vdots \\ \hat{y}_{t-p+1} \end{pmatrix} \quad \mu_t = \begin{pmatrix} Cz_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad F = \begin{pmatrix} A_1 & A_2 & \cdots & A_p \\ I_n & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I_n & 0 \end{pmatrix} \quad \zeta_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad K = \begin{pmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix} \tag{4.14.7}$$

In practice, we adapt Σ in K by setting $\Sigma_{ii} = 100$ whenever there is a condition on variable i at period t. This way the prior becomes uninformative and the condition set in (4.14.4) holds exactly in the posterior.

Finally, note that we can rewrite (4.14.4) in terms of $\hat{\gamma}_t$ instead of $\hat{y}_t$ by noting that $\hat{y}_t = Q\hat{\gamma}_t$, with $Q = (I_n \quad 0 \quad \cdots \quad 0)$ a $n \times np$ selection matrix that keeps only the first n rows of $\hat{\gamma}_t$:

$$\bar{y}_t = Q\hat{\gamma}_t + \epsilon_t \qquad \epsilon_t \sim N(0,\Omega_t) \qquad \Omega_t = diag(\omega_t) \tag{4.14.8}$$

Equations (4.14.8) and (4.14.6) respectively represent the observation and state equations of a state-space system in the state variable $\hat{\gamma}_t$. The first n rows of $\hat{\gamma}_t$ give the conditional forecasts $\hat{y}_t$, which represent the object of interest. Bayesian estimates for this state-space model can be obtained using standard approaches such as the Carter-Kohn algorithm[1].

Following, It is straightforward to define a Gibbs sampling algorithm for conditional forecasts:

algorithm 14.1: Gibbs sampling algorithm for conditional forecasts, basic approach

1. set the invariant matrices Z, $\bar{y}_t$ and Ω_t for $t = T+1, \cdots, T+h$.
2. 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. Recover $C^{(j)}, A_1^{(j)}, \cdots, A_p^{(j)}$ from $\beta^{(j)}$, and form $F^{(j)}$, $K^{(j)}$ and $\mu_t^{(j)}$ for $t = T+1, \cdots, T+h$.
3. obtain a sample $\hat{\gamma}_{T+1}^{(j)}, \cdots, \hat{\gamma}_{T+h}^{(j)}$ from a Bayesian state-space sampler such as the Carter-Kohn algorithm.
4. keep only the first n rows of $\hat{\gamma}_{T+1}^{(j)}, \cdots, \hat{\gamma}_{T+h}^{(j)}$ to obtain a sample $\hat{y}_{T+1}, \cdots, \hat{y}_{T+h}$ of conditional forecasts.
5. marginalize, that is, discard $\beta^{(j)}$ and $\Sigma^{(j)}$ and keep only the predictions $\hat{y}_{T+1}^{(j)}, \cdots, \hat{y}_{T+h}^{(j)}$.
6. repeat until the desired number of iterations is realised.

14.2 Conditional forecasts: a structural shock approach

The basic approach to conditional forecasts is agnostic about economic theory: it simply matches the dynamics of the model with the specified conditions. Sometimes, however, we want to assume that the conditions are generated by a subset of the structural shocks involved in the economy, providing more economic content to the exercise. The approach of conditional forecasts built on structural shocks has first

[1] Readers unfamiliar with state-space representations and Kalman filter methods should first read chapter K in Book 3 of the package.

been introduced by Waggoner and Zha (1999). It has then been amended by Andersson et al. (2010) to allow for density conditions, and we follow these lines here.

So, assume we have a VAR model and want to generate forecasts $\hat{y}_t$ for periods $t = T+1, \cdots, T+h$. From (4.13.9) (Wold theorem for structural impulse response function), the value of $\hat{y}_{T+h}$ can be expressed as:

$$\hat{y}_{T+h} = A(L)^{-1} C z_{T+h} + \Psi_0 \xi_{T+h} + \Psi_1 \xi_{T+h-1} + \Psi_2 \xi_{T+h-2} \cdots \tag{4.14.9}$$

Or:

$$\hat{y}_{T+h} = \underbrace{A(L)^{-1} C z_{T+h} + \sum_{i=0}^{\infty} \Psi_{h+i} \xi_{T-i}}_{\text{Forecast, absent future shocks}} + \underbrace{\sum_{j=1}^{h} \Psi_{h-j} \xi_{T+j}}_{\text{Impact of future shocks}} \tag{4.14.10}$$

The first two terms on the right-hand side of (4.14.10) represent the deterministic part of $\hat{y}_{T+h}$. They represent the forecast for y_{T+h} obtained with the observed data up to period T, when future shocks are unobserved. The second term on the right-hand side of (4.14.10) represents the contribution of future shocks to the realised value of y_{T+h}. Denoting the deterministic part by f_{T+h}, (4.14.10) rewrites:

$$\hat{y}_{T+h} = f_{T+h} + \sum_{j=1}^{h} \Psi_{h-j} \xi_{T+j} \tag{4.14.11}$$

Because f_{T+h} represents the forecast for y_{T+h} absent future shocks, it can easily be recovered numerically from the reduced-form VAR (4.11.1), by computing recursively $f_{T+1}, f_{T+2}, \cdots, f_{T+h}$, ignoring the shocks at each period. Formally, f_{T+h} obtains from:

$$\begin{aligned}
f_{T+1} &= C z_{T+1} + A_1 y_T + A_2 y_{T-1} + \cdots + A_p y_{T+1-p} \\
f_{T+2} &= C z_{T+2} + A_1 f_{T+1} + A_2 y_T + \cdots + A_p y_{T+2-p} \\
&\vdots \\
f_{T+h} &= C z_{T+h} + A_1 f_{T+h-1} + A_2 f_{T+h-2} + \cdots + A_p f_{T+h-p}
\end{aligned} \tag{4.14.12}$$

From (4.14.11), the prediction for $\hat{y}_{T+1}, \hat{y}_{T+2}, \cdots, \hat{y}_{T+h}$ can then write jointly as:

$$\begin{pmatrix} \hat{y}_{T+1} \\ \hat{y}_{T+2} \\ \hat{y}_{T+3} \\ \vdots \\ \hat{y}_{T+h} \end{pmatrix} = \begin{pmatrix} f_{T+1} \\ f_{T+2} \\ f_{T+3} \\ \vdots \\ f_{T+h} \end{pmatrix} + \begin{pmatrix} \Psi_0 & 0 & 0 & \cdots & 0 \\ \Psi_1 & \Psi_0 & 0 & \cdots & 0 \\ \Psi_2 & \Psi_1 & \Psi_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Psi_{h-1} & \Psi_{h-2} & \Psi_{h-3} & \cdots & \Psi_0 \end{pmatrix} \begin{pmatrix} \xi_{T+1} \\ \xi_{T+2} \\ \xi_{T+3} \\ \vdots \\ \xi_{T+h} \end{pmatrix} \tag{4.14.13}$$

Or, compactly:

$$\hat{y}_{T+1:T+h} = f_{T+1:T+h} + M\, \xi_{T+1:T+h} \qquad \xi_{T+1:T+h} \sim N(0, I_h \otimes \Gamma) \tag{4.14.14}$$

Hence, the unconditional forecasts have the following density:

$$\hat{y}_{T+1:T+h} \sim N(f_{T+1:T+h},\, M(I_h \otimes \Gamma)M') \tag{4.14.15}$$

Consider now imposing conditions on the forecasts. Precisely, we want to impose k conditions on the sequence of forecasts $\hat{y}_{T+1:T+h}$, where again each of these conditions can be expressed as:

$$\hat{y}_{i,t} \sim N(\bar{y}_{i,t}, \omega_{i,t}) \tag{4.14.16}$$

This can be represented in terms of $\hat{y}_{T+1:T+h}$ as:

$$R\, \hat{y}_{T+1:T+h} \sim N(\bar{y}, \Omega) \tag{4.14.17}$$

R is a $k \times nh$ selection matrix that takes a value of 1 for the values of $\hat{y}_{T+1:T+h}$ to which a condition applies. $\bar{y}$ is a k-dimensional vector of conditions, and $\Omega = diag(\omega)$, where ω is a k-dimensional vector of variances on the conditions. As usual, hard conditions can obtain by setting $\omega_i = 0$. The difficult part then consists in deriving the distribution of the shocks $\xi_{T+1:T+h}$ so that they satisfy at the same time the unconditional density (4.14.15) and the conditional density (4.14.17). After some work, it can be shown (book 2, p. 77) that the distribution of the constrained shocks is given by $\xi_{T+1:T+h} \sim N(\bar{\mu}, \bar{\Omega})$, with:

$$\bar{\mu} = D^*(\bar{y} - Rf_{T+1:T+h}) \qquad \bar{\Omega} = D^*\Omega D^{*\prime} + (I_{nh} - D^*D)(I_h \otimes \Gamma)(I_{nh} - D^*D) \tag{4.14.18}$$

where $D = RM$ is a $k \times nh$ matrix and D^* is the $nh \times k$ Moore-Penrose inverse of D such that $DD^* = I_k$. When $k \leq nh$, the matrix D^* is defined as:

$$D^* = D'(DD')^{-1} \tag{4.14.19}$$

Finally combining (4.14.18) with the unconditional forecast expression (4.14.14), one obtains that the distribution of the conditional forecasts is $\hat{y}_{T+1:T+h} \sim N(\hat{\mu}, \hat{\Omega})$, with (book 2, p. 78):

$$\hat{\mu} = f_{T+1:T+h} + MD^*(\bar{y} - Rf_{T+1:T+h}) \qquad \hat{\Omega} = M\left[D^*\Omega D^{*\prime} + (I_{nh} - D^*D)(I_h \otimes \Gamma)(I_{nh} - D^*D)\right] M' \tag{4.14.20}$$

The conditional forecast exercise then reduces to sampling from (4.14.20), which is straightforward. But this methodology permits in fact to do even better. Often, we want the conditions to be generated by a subset of the structural shocks only. For instance, we may assume that the conditions applicable on the interest rate obtain only from monetary policy shocks. Antolin-Diaz et al. (2018) notice that this can be achieved by setting the condition:

$$P\xi_{T+1:T+h} \sim N(0, \Gamma_{nd}) \tag{4.14.21}$$

where P is a $m \times nh$ selection matrix formed by ones and zeros that takes a value of 1 to select the m shocks that do *not* drive the conditions. Γ_{nd} is a $m \times m$ diagonal matrix that select the entries of Γ corresponding to the variances of the non-driving shocks[2]. This way (4.14.21) constrains the non-driving shocks to keep their unconditional distributions, permitting only the remaining driving shocks to generate the conditions. Now, note that the unconditional forecasts (4.14.14) can rewrite

$$M^{-1}\hat{y}_{T+1:T+h} = M^{-1}f_{T+1:T+h} + \xi_{T+1:T+h} \qquad \xi_{T+1:T+h} \sim N(0, I_h \otimes \Gamma) \tag{4.14.22}$$

Pre-multiplying by P:

$$Q\hat{y}_{T+1:T+h} = Qf_{T+1:T+h} + P\xi_{T+1:T+h} \qquad \xi_{T+1:T+h} \sim N(0, I_h \otimes \Gamma) \qquad Q = PM^{-1} \tag{4.14.23}$$

And this eventually implies the restrictions:

$$Q\,\hat{y}_{T+1:T+h} \sim N(Qf_{T+1:T+h}, \Gamma_{nd}) \tag{4.14.24}$$

To obtain conditional forecasts that satisfy at the same time the conditions (4.14.17) and the restrictions on driving shocks (4.14.24), we simply stack them to obtain:

$$Z\,\hat{y}_{T+1:T+h} \sim N(g_{T+1:T+h}, \Xi) \tag{4.14.25}$$

with:

$$Z = \begin{pmatrix} R \\ Q \end{pmatrix} \qquad g_{T+1:T+h} = \begin{pmatrix} \bar{y} \\ Qf_{T+1:T+h} \end{pmatrix} \qquad \Xi = \begin{pmatrix} \Omega & 0 \\ 0 & \Gamma_{nd} \end{pmatrix} \tag{4.14.26}$$

Equation (4.14.25) is similar to (4.14.17), so that adding shock restrictions can be reduced to a regular conditional forecast settings. Following, the conditional forecast distribution still obtains directly from

[2] Specifically, $\Gamma_{nd} = diag(P\gamma)$, where γ is the vector obtained from the main diagonal of $I_h \otimes \Gamma$.

(4.14.20), replacing R, $\bar{y}$ and Ω with Z, $g_{T+1:T+h}$ and Ξ.

We can then propose the following Gibbs sampling algorithm for conditional forecasts:

algorithm 14.2: Gibbs sampling algorithm for conditional forecasts, structural shocks approach

1. set the invariant matrices R, $\bar{y}$, Ω, and P if shock restrictions apply.
2. 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. Recover $C^{(j)}, A_1^{(j)}, \cdots, A_p^{(j)}$ from $\beta^{(j)}$.
3. form $f_{T+1:T+h}$ from $C^{(j)}, A_1^{(j)}, \cdots, A_p^{(j)}$.
4. form $\Psi_0, \Psi_1, \cdots, \Psi_{h-1}$ from $\beta^{(j)}$, then construct M.
5. if conditions on shocks apply, compute also Q, Z, $g_{T+1:T+h}$, Γ_{nd} and Ξ.
6. compute D, $\hat{D}$, D^*, $\hat{\mu}$ and $\hat{\Omega}$.
7. draw $\hat{y}^{(j)}_{T+1:T+h}$ from $\hat{y}^{(j)}_{T+1:T+h} \sim N(\hat{\mu}, \hat{\Omega})$.
8. marginalize, that is, discard $\beta^{(j)}$, $\Sigma^{(j)}$ and $\Gamma^{(j)}$ and keep only the predictions $\hat{y}^{(j)}_{T+1}, \cdots, \hat{y}^{(j)}_{T+h}$.
9. repeat until the desired number of iterations is realised.

14.3 Structural identification by sign and zero restrictions

Section 13.2 introduced the notion of structural identification, along with the common approach of Cholesky factorization. Another popular approach to structural identification is the **sign restrictions** methodology introduced by Arias et al. (2018). In this approach, the structural identification is generated by restrictions on the signs of the structural impulse response functions, making sure that they are consistent with economic theory.

Consider the general structural VAR introduced in definition 13.1:

$$H_0 y_t = G z_t + H_1 y_{t-1} + \cdots + H_p y_{t-p} + \xi_t \qquad \xi_t \sim N(0, I_n) \qquad t = 1, \cdots, T \tag{4.14.27}$$

where for simplicity it is assumed that the structural shocks have unit variance. The aim of the sign restriction exercise consists in verifying whether the structural impulse response functions $\Psi_0, \Psi_1, \Psi_2, \cdots$ produced by the SVAR (4.14.27) satisfy a set of restrictions specified by the user.

In general, we may want to test for three kinds of restrictions. Pure **sign restrictions** test for the sign of the structural response of a variable to a given structural shock, at some specific horizon. That is, it tests whether $\psi^h_{ij} > 0$ (alternatively $\psi^h_{ij} < 0$) for variable i, structural shocks j and horizon h. **magnitude restrictions** test whether the magnitude of the structural response to one structural shock is larger (in absolute value) than the magnitude of the response to some other shock. That is, it tests whether $|\psi^h_{ij}| > |\psi^h_{ik}|$, for two structrual shocks j and k. Finally, **zero restrictions** test whether the impact of some impulse response was null. That is, it tests whether $\psi^h_{ij} = 0$.

Testing for a set of k restrictions can be done by the way of an impulse-response function matrix $f(\Psi)$ that stacks the structural impulse response functions for all the horizons at which restrictions apply, and pairs of selection matrices e_i and s_i, for $i = 1, \cdots, k$. To make things more concrete, consider a simple VAR with two variables and two structural shocks. We implement $k = 3$ restrictions covering horizons 0 and 1 (impact and one period after). Then $f(\Psi)$ is given by:

$$f(\Psi) = \begin{pmatrix} \Psi_0 \\ \Psi_1 \end{pmatrix} = \begin{pmatrix} \psi_{11}^0 & \psi_{12}^0 \\ \psi_{21}^0 & \psi_{22}^0 \\ \psi_{11}^1 & \psi_{12}^1 \\ \psi_{21}^1 & \psi_{22}^1 \end{pmatrix} \tag{4.14.28}$$

For the first restriction, we want the response of variable 1 to structural shock 2 at horizon 1 to be positive, that is, $\psi_{12}^1 > 0$. The restriction will hold if:

$$e_1 \times f(\Psi) \times s_1 > 0 \quad \text{or} \quad \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \psi_{11}^0 & \psi_{12}^0 \\ \psi_{21}^0 & \psi_{22}^0 \\ \psi_{11}^1 & \psi_{12}^1 \\ \psi_{21}^1 & \psi_{22}^1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} > 0 \quad \text{or} \quad \psi_{12}^1 > 0 \tag{4.14.29}$$

For the second restriction, we want the response of variable 2 to shock 1 to be smaller in magnitude than the response to shock 2 at horizon 1. That is, we want $|\psi_{21}^1| < |\psi_{22}^1|$. This will hold if:

$$e_2 \times |f(\Psi)| \times s_2 > 0 \quad \text{or} \quad \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} |\psi_{11}^0| & |\psi_{12}^0| \\ |\psi_{21}^0| & |\psi_{22}^0| \\ |\psi_{11}^1| & |\psi_{12}^1| \\ |\psi_{21}^1| & |\psi_{22}^1| \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} > 0 \quad \text{or} \quad |\psi_{21}^1| < |\psi_{22}^1| \tag{4.14.30}$$

For the final restriction, we want the response of variable 2 to shock 1 to be null at impact. That is, we want $\psi_{21}^0 = 0$. The restriction will hold if:

$$e_3 \times f(\Psi) \times s_3 = 0 \quad \text{or} \quad \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \psi_{11}^0 & \psi_{12}^0 \\ \psi_{21}^0 & \psi_{22}^0 \\ \psi_{11}^1 & \psi_{12}^1 \\ \psi_{21}^1 & \psi_{22}^1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 0 \quad \text{or} \quad \psi_{21}^0 = 0 \tag{4.14.31}$$

The above procedure makes it trivial to verify whether a specific structural VAR satisfies a given set of restrictions. Assume for now that we deal with sign restrictions only. It will become clear shorlty why the zero restrictions represent a special case. Then we can propose the following Gibbs sampling algorithm:

algorithm 14.3: Gibbs sampling algorithm for sign and magnitude restrictions, SVAR parameterization

1. at iteration j, draw $H_0^{(j)}$, $G^{(j)}$ and $H_1^{(j)}, \cdots, H_p^{(j)}$ from their posterior distributions.
2. generate $\Psi_0, \Psi_1, \Psi_2, \cdots$, and $f(\Psi)$.
3. for $i = 1, \cdots, k$, test the restriction by verifying if $e_i \times f(\Psi) \times s_i > 0$ (for sign restrictions), or $e_i \times |f(\Psi)| \times s_i > 0$ (for magnitude restrictions).
4. keep the draw if all k restrictions are satisfied; else, reject and go back to step 1.
5. repeat until the desired number of sucessful iterations is realised.

The problem with algorithm 14.3 is that in practical applications it is usually not possible to sample directly from the SVAR parameters $H_0, G, H_1, \cdots, H_p$. Rather, we can sample from the reduced-form parameters B and Σ. Therefore, we need a mapping from the reduced-form parameters to the structural parameters. To do so, Arias et al. (2018) propose a parameterization called the **orthogonal reduced-form parameterization**, given by:

$$y_t = Cz_t + A_1 y_{t-1} + \cdots + A_p y_{t-p} + h(\Sigma)\, Q\, \xi_t \qquad \xi_t \sim N(0, I) \qquad t = 1, \cdots, T \tag{4.14.32}$$

The matrix $h(\Sigma)$ is any decomposition of Σ such that $h(\Sigma)h(\Sigma)' = \Sigma$. In practice, we take $h(\Sigma)$ to be the Cholesky factor, though any differentiable decomposition would do. Q is a $n \times n$ orthogonal matrix such that $QQ' = I_n$. It plays the role of a rotation matrix applied to the original decomposition $h(\Sigma)$. It is easy to see that this formulation is equivalent to the regular reduced-form VAR (4.11.1). Indeed, we can define $\varepsilon_t = h(\Sigma)\ Q\ \xi_t$ and obtain:

$$Var(\varepsilon_t) = Var(h(\Sigma)\ Q\ \xi_t) = h(\Sigma)\ Q\ I_n\ Q'\ h(\Sigma)' = h(\Sigma)\ Q\ Q'\ h(\Sigma)' = h(\Sigma)\ h(\Sigma)' = \Sigma \tag{4.14.33}$$

It is also easy to notice that $H = h(\Sigma)\ Q$ is the structural decomposition matrix of the model since $HH' = \Sigma$. Pre-multiplying the orthogonal reduced-for VAR (4.14.32) by $H_0 = H^{-1}$, one recovers the structural VAR (4.14.27). Conditional on the orthogonal reduced-form parameters B, Σ and Q, one can thus recover the SVAR parameters $H_0, G, H_1, \cdots, H_p$, and test for the sign restrictions.

It remains to obtain random draws for the orthogonal matrix Q. Ideally, the draws would obtain from a uniform distribution since we are agnostic about which values Q should take. Arias et al. (2018) propose a simple method to obtain Q:

algorithm 14.4: Sampling of orthogonal matrix Q from uniform distribution

1. generate a $n \times n$ matrix X with each element having an independent standard normal distribution.
2. compute $QR = X$, the QR decomposition of X.
3. if needed, normalize Q and R so that the diagonal of R has only positive entries; then Q has the uniform ditribution over $\mathcal{O}(n)$, the set of $n \times n$ orthogonal matrices.

It is then possible to define the following Gibbs sampling algorithm for restrictions:

algorithm 14.5: Gibbs sampling algorithm for sign and magnitude restrictions, orthogonal reduced-form parameterization

1. at iteration j, draw the reduced-form parameters $B^{(j)}$ and $\Sigma^{(j)}$ from their posterior distributions.
2. from $B^{(j)}$, obtain the impulse-response function $\Phi_0, \Phi_1, \Phi_2 \cdots$.
3. from $\Sigma^{(j)}$, obtain the decomposition $h(\Sigma)$.
4. generate an orthogonal matrix Q from algorithm 14.4.
5. create a candidate structural decomposition matrix $H = h(\Sigma)\ Q$.
6. generate the structural impulse response function $\Psi_0, \Psi_1, \Psi_2 \cdots$ from $\Psi_i = \Phi_i\ H$; generate the matrix $f(\Psi)$.
7. for $i = 1, \cdots, k$, test the restriction by verifying if $e_i \times f(\Psi) \times s_i > 0$ (for sign restrictions), or $e_i \times |f(\Psi)| \times s_i > 0$ (for magnitude restrictions).
8. keep H if all k restrictions are satisfied; else, reject and go back to step 1.
9. repeat until the desired number of sucessful iterations is realised.

Consider now adding zero restrictions. Intuitively, we would like to use algorithm 14.5 for zero restrictions as well, but this is not possible. The problem is that the the set of random matrices Q satifying the zero restrictions has measure zero. That is, the probability to obtain by chance a random rotation matrix Q that will exactly satisfy the zero restrictions is null. For this reason, the matrix Q must be constructed column by column so as to satisfy the restrictions. Arias et al. (2018) propose the following algorithm:

algorithm 14.6: Construction of matrix Q satisfying the zero restrictions

for $j = 1, 2, \cdots, n$:

1. construct the matrix Z_j that stacks the e_i vectors related to zero restrictions on structural shock j. If there are no zero restrictions on shock j, define Z_j to be the empty matrix.
2. construct the matrix Q'_{j-1}, where Q_0 is the empty matrix, and otherwise $Q_{j-1} = [q_1 \; q_2 \; \cdots \; q_{j-1}]$ denotes the set of columns of Q previously created.
3. construct the matrix $R_j = \begin{pmatrix} Z_j \times f(\Psi) \\ Q'_{j-1} \end{pmatrix}$.
4. draw a random vector x_j from a standard normal distribution on $\mathbb{R}^n$.
5. if R_j is empty, define $q_j = x_j / \|x_j\|$.
6. if R_j is non-empty, find a matrix N_j whose columns form a non-zero orthonormal basis for the nullspace of R_j; then define $q_j = N_j(N'_j x_j / \|N'_j x_j\|)$.
7. set q_j as column j of Q.

The methodology implies that no more than $(n-j)$ zero restrictions can be set on structural shock j, otherwise the matrix Q is not identified[3]. As the ordering of variables does not matter with sign restrictions, one can play on the ordering to cope with the desired number of zero restrictions on the different variables.

With algorithm 14.6, it is possible to develop a Gibbs sampling procedure for the general case of sign, magnitude and zero restrictions:

algorithm 14.7: Gibbs sampling algorithm for sign and magnitude restrictions, orthogonal reduced-form parameterization

1. at iteration j, draw the reduced-form parameters $B^{(j)}$ and $\Sigma^{(j)}$ from their posterior distributions.
2. from $B^{(j)}$, obtain the impulse-response function $\Phi_0, \Phi_1, \Phi_2 \cdots$.
3. from $\Sigma^{(j)}$, obtain the decomposition $h(\Sigma)$. Get a preliminary matrix $f(\tilde{\Psi})$ from $\tilde{\Psi}_i = \Phi_i \, h(\Sigma)$.
4. generate an orthogonal matrix Q from algorithm 14.6, using $f(\tilde{\Psi})$.
5. create a candidate structural decomposition matrix $H = h(\Sigma) \, Q$.
6. generate the structural impulse response function $\Psi_0, \Psi_1, \Psi_2 \cdots$ from $\Psi_i = \Phi_i \, H$; generate the matrix $f(\Psi)$.
7. for $i = 1, \cdots, k$, test the restriction by verifying if $e_i \times f(\Psi) \times s_i > 0$ (for sign restrictions), or $e_i \times |f(\Psi)| \times s_i > 0$ (for magnitude restrictions). Zero restrictions need not be verified since they are satisfied by construction.
8. keep H if all k restrictions are satisfied; else, reject and go back to step 1.
9. repeat until the desired number of sucessful iterations is realised.

This concludes the presentation of the sign restrictions methodology. A final remark applies: in case of pure sign restrictions, algorithm 14.5 is equivalent to sampling directly from the posterior distribution of the SVAR parameters. When zero restrictions are involved, however, algorithm 14.7 produces structural parameter draws from a different distribution. To remedy this problem, Arias et al. (2018) propose an importance sampling procedure. We do not follow this line, for two reasons at least. First, in practice,

[3] If ones tries to impose more than $(n-j)$ zero restrictions on structural shock j, then the basis N_j of the nullspace of R_j can only be the trivial zero vector. This in turns implies that q_j is also a zero vector so that Q cannot be orthogonal.

using or not the importance sampling procedure generates similar distributions. Second, the importance sampling procedure is computationally expensive and may render estimation intractable. Also, it requires the computation of numerical derivatives, making it prone to numerical error and instability. In every respect, it seems simpler and safer to apply algorithm 14.7 directly.

14.4 Structural identification by narrative sign restrictions

Antolin-Diaz and Rubio-Ramírez (2018) propose to extend the class of sign restriction methodologies to structural shocks and historical decomposition. They call this new category of restrictions the **narrative sign restrictions**. The overall identification procedure is similar to that of traditional sign restrictions and only requires some adaptation to account for the alternative applications to which the restrictions apply.

Consider first **structural shock restrictions**. These restrictions apply either to the sign of the j^{th} structural shock at some sample period t (e.g. $\xi_{i,t} > 0$), or to the relative magnitudes of shocks at period t (e.g. $|\xi_{i,t}| > |\xi_{j,t}|$). Similar to regular sign restrictions, we can test for the restrictions with a structural shock matrix $f(\xi)$ stacking the relevant vectors of in-sample structural shocks on which restrictions apply, and pairs of selection matrices e_i and s_i, for $i = 1, \cdots, k$ restrictions.

Consider again the case of a simple VAR with two variables and two structural shocks. We implement $k = 2$ shock restrictions covering sample periods $t = 50$ and 51. Then $f(\xi)$ is given by:

$$f(\xi) = \begin{pmatrix} \xi'_{50} \\ \xi'_{51} \end{pmatrix} = \begin{pmatrix} \xi_{1,50} & \xi_{2,50} \\ \xi_{1,51} & \xi_{2,51} \end{pmatrix} \tag{4.14.34}$$

For the first restriction, we want the first structural shock to be negative at sample period $t = 51$, that is, $\xi_{1,51} < 0$. The restriction will hold if:

$$e_1 \times f(\xi) \times s_1 > 0 \quad \text{or} \quad \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_{1,50} & \xi_{2,50} \\ \xi_{1,51} & \xi_{2,51} \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix} > 0 \quad \text{or} \quad \xi_{1,51} < 0 \tag{4.14.35}$$

The second restriction considers that the first structural shock at period $t = 50$ is larger in magnitude than the second structural shock, that is, $|\xi_{1,50}| > |\xi_{2,50}|$. The restriction will hold if:

$$e_1 \times |f(\xi)| \times s_1 > 0 \quad \text{or} \quad \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} |\xi_{1,50}| & |\xi_{2,50}| \\ |\xi_{1,51}| & |\xi_{2,51}| \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} > 0 \quad \text{or} \quad |\xi_{1,50}| > |\xi_{2,50}| \tag{4.14.36}$$

Consider now the case of **historical decomposition restrictions**. Remember from (4.13.35) and (4.13.36) that the historical decomposition of sample observation $y_{i,t}$ is given by:

$$y_{i,t} = d_{i,t} + h_{i1,t} + h_{i2,t} + \cdots + h_{in,t} \qquad h_{ik,t} = \sum_{j=0}^{t-1} \psi_{ik}^{j} \xi_{k,t-j} \tag{4.14.37}$$

Each $h_{ij,t}$ represents the historical contribution of shock j to the value of variable i at sample period t. Again, two types of restrictions apply. We can implement restrictions on the sign of the historical contribution of structural shock j on variable i at some sample period t (e.g. $h_{ij,t} > 0$). Or we can apply restrictions on the relative magnitudes of the historical contributions of structural shocks j and k on variable i at some sample period t (e.g. $|h_{ij,t}| > |h_{ik,t}|$). The restrictions can be checked with a historical decomposition matrix $f(h)$ stacking the relevant vectors of in-sample historical contributions on which restrictions apply, and pairs of selection matrices e_i and s_i, for $i = 1, \cdots, k$ restrictions.

Consider again the case of a simple VAR with two variables and two structural shocks. We implement $k = 2$ historical restrictions covering sample periods $t = 50$ for variable 1 and period $t = 51$ for variable 2. Then $f(h)$ is given by:

$$f(h) = \begin{pmatrix} h'_{1,50} \\ h'_{2,51} \end{pmatrix} = \begin{pmatrix} h_{11,50} & h_{12,50} \\ h_{21,51} & h_{22,51} \end{pmatrix} \tag{4.14.38}$$

For the first restriction, we want the historical contribution of structural shock 1 to variable 1 at period $t = 50$ to be positive, that is, $h_{11,50} > 0$. The restriction will hold if:

$$e_1 \times f(h) \times s_1 > 0 \quad \text{or} \quad \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} h_{11,50} & h_{12,50} \\ h_{21,51} & h_{22,51} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} > 0 \quad \text{or} \quad h_{11,50} > 0 \tag{4.14.39}$$

The second restriction considers that the historical contribution of structural shock 2 to variable 2 at period $t = 51$ is larger than that of structural shock 1, that is, $|h_{22,51}| > |h_{21,51}|$. The restriction will hold if:

$$e_1 \times |f(h)| \times s_1 > 0 \quad \text{or} \quad \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} |h_{11,50}| & |h_{12,50}| \\ |h_{21,51}| & |h_{22,51}| \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} > 0 \quad \text{or} \quad |h_{22,51}| > |h_{21,51}| \tag{4.14.40}$$

With these elements, we can propose a simple Gibbs sampling procedure for the narrative sign restrictions:

algorithm 14.8: Gibbs sampling algorithm for narrative sign restrictions

1. at iteration j, draw the reduced-form parameters $B^{(j)}$ and $\Sigma^{(j)}$ from their posterior distributions.
2. from $B^{(j)}$, obtain the reduced-form residuals $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_T$.
3. from $B^{(j)}$, obtain the impulse-response function $\Phi_0, \Phi_1, \Phi_2 \cdots$.
4. from $\Sigma^{(j)}$, obtain the decomposition $h(\Sigma)$.
5. generate an orthogonal matrix Q from $QR = X$, with X a $n \times n$ matrix with each element having an independent standard normal distribution, and the diagonal of R normalized to be positive.
6. create a candidate structural decomposition matrix $H = h(\Sigma)\, Q$.
7. generate the structural shocks $\xi_1, \xi_2, \cdots, \xi_T$ from the $\xi_t = H^{-1}\varepsilon_t$. Generate $f(\xi)$.
8. generate the structural impulse response function $\Psi_0, \Psi_1, \Psi_2 \cdots$ from $\Psi_i = \Phi_i\, H$.
9. generate the historical decomposition $h_{ij,t}$, for each triplet i, j, t on which some restriction applies. Generate $f(h)$.
10. for $i = 1, \cdots, k$, test the restriction by verifying if $e_i \times f(\xi) \times s_i > 0$ (for sign restrictions on shocks), $e_i \times |f(\xi)| \times s_i > 0$ (for magnitude restrictions on shocks), $e_i \times f(h) \times s_i > 0$ (for historical sign restrictions), or $e_i \times |f(h)| \times s_i > 0$ (for historical magnitude restrictions).
11. keep H if all k restrictions are satisfied; else, reject and go back to step 1.
12. repeat until the desired number of sucessful iterations is realised.

14.5 Structural identification by proxy-SVAR

The approaches discussed so far identify the structural VAR model by imposing some restrictions on the sign of the shocks, impulse response functions, or historical decomposition. An alternative approach consists in identifying the structural VAR by using external instruments known as **proxy**, assuming that these instruments carry some information about the structural shocks to be identified. While there exist a number of contributions on the frequentist side, few Bayesian methodologies have been developed. Caldara and Herbst (2019) propose a Metropolis-Hastings procedure that relies on a single proxy. A more general approach is developed by Arias et al. (2021), and the presentation in this section follows the same line.

Consider a general structural VAR of the form:

$$H_0 y_t = G z_t + H_1 y_{t-1} + \cdots + H_p y_{t-p} + \xi_t \qquad \xi_t \sim N(0, I_n) \qquad t = 1, \cdots, T \tag{4.14.41}$$

where $y_t, y_{t-1}, \cdots, y_{t-p}$ and ξ_t are n-dimensional vectors of observations and structural shocks, and $H_0, H_1, \cdots, H_p$ are $n \times n$ matrices of coefficients. Consider adding a vector r_t containing h external instruments or proxy to the structural VAR. Then model (4.14.41) can rewrite jointly with r_t as:

$$\begin{pmatrix} H_0 & 0_{n\times h} \\ \Gamma_{0,1} & \Gamma_{0,2} \end{pmatrix} \begin{pmatrix} y_t \\ r_t \end{pmatrix} = \begin{pmatrix} G \\ F \end{pmatrix} (z_t) + \begin{pmatrix} H_1 & 0_{n\times h} \\ \Gamma_{1,1} & \Gamma_{1,2} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ r_{t-1} \end{pmatrix} + \cdots + \begin{pmatrix} H_p & 0_{n\times h} \\ \Gamma_{p,1} & \Gamma_{p,2} \end{pmatrix} \begin{pmatrix} y_{t-p} \\ r_{t-p} \end{pmatrix} + \begin{pmatrix} \xi_t \\ \nu_t \end{pmatrix} \tag{4.14.42}$$

where $\Gamma_{i,1}$ and $\Gamma_{i,2}$ respectively denote matrices of coefficients of dimension $h \times n$ and $h \times h$, for $i = 0, 1, \cdots, p$. ν_t is a h-dimensional vector of structural shock specific to the instrument and uncorrelated with ξ_t, with $\nu_t \sim N(0, I_h)$. Note that the original SVAR (4.14.41) implies blocks of zeros in the coefficient matrices of the augmented SVAR (4.14.42). These are known as the **block restrictions** of the proxy SVAR.

Model (4.14.42) can be written in compact form as:

$$\bar{H}_0 \bar{y}_t = \bar{G} z_t + \bar{H}_1 \bar{y}_{t-1} + \cdots + \bar{H}_p \bar{y}_{t-p} + \bar{\xi}_t \qquad \bar{\xi}_t \sim N(0, I_{\bar{n}}) \qquad \bar{n} = n + h \tag{4.14.43}$$

Stacking the regressors and coefficient matrices, (4.14.43) can rewrite:

$$\bar{H}_0 \bar{y}_t = \bar{H}_+ \bar{x}_t + \bar{\xi}_t \qquad \bar{H}_+ = \begin{pmatrix} \bar{G} & \bar{H}_1 & \cdots & \bar{H}_p \end{pmatrix} \qquad \bar{x}_t = \begin{pmatrix} z_t' & \bar{y}_{t-1}' & \cdots & \bar{y}_{t-p}' \end{pmatrix}' \tag{4.14.44}$$

The h external instruments in r_t are assumed to be correlated with h structural shocks in ξ_t, and to be uncorrelated with the other shocks. The first assumption is known as the **relevance conditions**, and states that the proxy are expected to carry information about the structural shocks to which they are related. The second assumption is known as the **exogeneity restrictions**, and states that the proxy don't carry information beyond the structural shocks they represent. Without loss of generality, let the h proxy be related to the last h structural shocks, and uncorrelated with the first $n - h$ elements of ξ_t.

Consider $\bar{H}_0$. It is easy to show (book 2, p. 78) that its inverse $\bar{H}_0^{-1}$ is given by:

$$\bar{H}_0^{-1} = \begin{pmatrix} H_0^{-1} & 0_{n\times h} \\ -\Gamma_{0,2}^{-1}\, \Gamma_{0,1}\, H_0^{-1} & \Gamma_{0,2}^{-1} \end{pmatrix} \tag{4.14.45}$$

Using (4.14.45) with the relevance and exogeneity restrictions, it can then be shown (book 2, p. 79) that:

$$\mathbb{E}(r_t \xi_t') = [0_{h\times(n-h)} \quad V] = -\Gamma_{0,2}^{-1}\, \Gamma_{0,1}\, H_0^{-1} \tag{4.14.46}$$

The first equality in (4.14.46) reflects the fact that the first $n-h$ structural shocks are orthogonal to the proxy variables (exogeneity restrictions), while the last h structural shocks are correlated with the h proxies through the covariance matrix V (relevance conditions). The second equality shows that identifying a proxy-SVAR is realized through zero restrictions on the structural parameters $\Gamma_{0,1}$, $\Gamma_{0,2}$ and H_0^{-1} (the exogeneity restrictions). Note in particular that (4.14.46) implies zero restrictions on the columns of $H = H_0^{-1}$, the structural identification matrix of the SVAR (4.14.41).

Intuitively, we would like to apply the methodology of Arias et al. (2018) on sign and zero restrictions to the proxy SVARs: estimate the Bayesian VAR under the orthogonal reduced-form parameterization, then sample from the distribution over the structural parameterization of the proxy-SVAR conditional on the block and exogeneity restrictions (4.14.42) and (4.14.45). Unfortunately, this is not possible because the implied number of zero restrictions is too large. For this reason, Arias et al. (2021) propose an alternative parameterization called the **orthogonal triangular-block parameterization**.

Concretely, the orthogonal triangular-block parameterization works as follows. Let $\bar{\Lambda}_0$ be a $\bar{n} \times \bar{n}$ matrix restricted to be lower-triangular with positive diagonal. Let $\bar{\Lambda}_+$ be a $\bar{n} \times \bar{k}$ matrix, where $\bar{k} = m + \bar{n}p$ and m denotes as usual the number of exogenous regressors in z_t. $\bar{\Lambda}_+$ is defined as $\bar{\Lambda}_+ = (D \quad \Lambda_1 \quad \cdots \quad \Lambda_p)$, where D is a $\bar{n} \times m$ matrix and each Λ_i is $\bar{n} \times \bar{n}$ and restricted so that the upper right $n \times h$ block is zero, similar to $\bar{H}_i$. Finally let Q_1 and Q_2 respectively denote $n \times n$ and $h \times h$ orthogonal matrices, and let $Q = diag(Q_1, Q_2)$ be a $\bar{n} \times \bar{n}$ block-diagonal orthogonal matrix. $\bar{\Lambda}_0$, $\bar{\Lambda}_+$ and Q together define the orthogonal triangular-block parameterization, which writes as:

$$\bar{\Lambda}_0 \, \bar{y}_t = \bar{\Lambda}_+ \, \bar{x}_t + \bar{u}_t \qquad \bar{u}_t = Q' \, \bar{\xi}_t \tag{4.14.47}$$

It is then easy to see that we can obtain parameters $\bar{H}_0$ and $\bar{H}_+$ that satisfy the block restrictions and thus satisfy (4.14.44) by applying the mapping:

$$\bar{H}_0 = Q \, \bar{\Lambda}_0 \qquad \bar{H}_+ = Q \, \bar{\Lambda}_+ \tag{4.14.48}$$

The orthogonal triangular-block representation (4.14.47) is similar to the proxy-SVAR representation (4.14.44) up to a pre-multiplication by Q, which thus plays here the role of a rotation matrix mapping the two representations.

The first step of the orthogonal triangular-block approach then consists in obtaining values for $\bar{\Lambda}_0$ and $\bar{\Lambda}_+$. Arias et al. (2021) propose to sample these parameters from a restricted **normal-generalized-normal distribution**, which is a conjugate posterior distribution satisfying the block restrictions. The procedure is involving, so we will mainly outline the procedure as a cookbook. First, define the stacked data matrix for the SVAR data as:

$$\bar{Y} = \begin{pmatrix} \bar{y}_1 & \bar{y}_2 & \cdots & \bar{y}_T \end{pmatrix}' \qquad \bar{X} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_T \end{pmatrix}' \tag{4.14.49}$$

where $\bar{Y}$ and $\bar{X}$ are of respective dimensions $T \times \bar{n}$ and $T \times \bar{k}$. The posterior distribution involves four posterior parameters $\bar{\alpha}$, $\bar{W}$, $\bar{B}$ and $\bar{S}$, defined as:

$$\bar{\alpha} = \alpha + T \qquad \bar{W} = (W^{-1} + \bar{X}'\bar{X})^{-1} \qquad \bar{B} = \bar{W}(W^{-1}B + \bar{X}'\bar{Y}) \qquad \bar{S} = S + \bar{Y}'\bar{Y} + B'W^{-1}B - \bar{B}'\bar{W}^{-1}\bar{B} \tag{4.14.50}$$

$\bar{\alpha}$ is a scalar-valued shape parameter and $\bar{S}$ is a $\bar{n} \times \bar{n}$ scale matrix. $\bar{W}$ and $\bar{B}$ respectively denote $\bar{k} \times \bar{k}$ and $\bar{k} \times \bar{n}$ covariance and location matrices. Arias et al. (2021) suggest to set the prior parameters α, W, B and S to $\alpha = \bar{n}$, $W^{-1} = 0$, $B = 0$ and $S = 0$. However, more sensitive values can be used, such as the prior values α, W, B and S used for the normal-Wishart as described in section 11.3.

The difficult part consists in sampling $\bar{\Lambda}_0$ and $\bar{\Lambda}_+$ that satisfy the orthogonal triangular-block representation. This can be done with the Gibbs sampling algorithm of Waggoner and Zha (2003), which provides a method to sample SVAR parameters subject to certain class of linear restrictions. Applied to the proxy SVAR parameters, this yields the following algorithm:

algorithm 14.9: Sampling of SVAR parameters $\bar{\Lambda}_0$ and $\bar{\Lambda}_+$

For $j = 1, \cdots, \bar{n}$:

1. generate the matrices U_j and V_j. U_j is defined as the first j columns of $I_{\bar{n}}$. For $j = 1, \cdots, n$, V_j is block diagonal with $p+1$ block; the first block (for the exogenous regressors) is I_m, and the other p blocks (one for each lag of the endogenous) are $\bar{n} \times n$ matrices made of the first n columns of $I_{\bar{n}}$. For $j = n+1, \cdots, \bar{n}$, V_j is defined as $I_{\bar{k}}$.

2. define the matrices:
$H_j = (V_j' \, \bar{W}^{-1} \, V_j)^{-1}$
$P_j = H_j \, V_j' \, \bar{W}^{-1} \, \bar{B} \, U_j$
$Q_j = \bar{\alpha} \, (U_j' \, \bar{S} \, U_j + U_j' \, \bar{B}' \, \bar{W}^{-1} \, \bar{B} \, U_j - P_j' \, H_j^{-1} \, P_j)^{-1}$

3. find a non-zero vector z that is orthogonal to all rows of $\bar{\Lambda}_0$, save row j. Use the values of $\bar{\Lambda}_0$ obtained from the latest Gibbs sampler iteration.

4. define the vector $w_1 = F_j' \, U_j' \, z / \| F_j' \, U_j' \, z \|$, where F_j is the Cholesky factor of Q_j.

5. build the vectors $w_2, \cdots, w_j$ recursively so that they form an orthonormal basis for $\mathbb{R}^j$. To do so, denote $w_1' = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,j} \end{pmatrix}$. Then for $i = 2, \cdots, j$, define:
$w_i' = \begin{pmatrix} w_{1,1} w_{1,i} & \cdots & w_{1,i-1} w_{1,i} & -c_{i-1} & 0 & \cdots & 0 \end{pmatrix} / \sqrt{c_{i-1} c_i}$ with $c_i = \sum_{k=1}^{i} w_{1,k}^2$

6. define the vector $s = \begin{pmatrix} s_1 & \cdots & s_{\bar{\alpha}+1} \end{pmatrix}'$, where each s_i is drawn from $s_i \sim N(0, 1/\bar{\alpha})$. Then define $r = s's$, and finally assign $\beta_1 = \sqrt{(r)}$ or $\beta_1 = -\sqrt{(r)}$, each with probability one-half.

7. draw β_i from $\beta_i \sim N(0, 1/\bar{\alpha})$, for $i = 2, \cdots, j$.

8. define $\gamma_{0,j} = F_j \sum_{i=1}^{j} \beta_i \, w_i$. If needed, multiply by -1 so that entry j of $\gamma_{0,j}$ is positive; this ensures a positive diagonal for $\bar{\Lambda}_0$.

9. draw $\gamma_{+,j}$ from $\gamma_{+,j} \sim N(P_j \, \gamma_{0,j}, H_j)$.

10. generate $\lambda_{0,j}$, the j^{th} row of $\bar{\Lambda}_0$, and $\lambda_{+,j}$, the j^{th} row of $\bar{\Lambda}_+$, from:
$\lambda_{0,j} = U_j \, \gamma_{0,j}$ $\lambda_{+,j} = V_j \, \gamma_{+,j}$
Update $\bar{\Lambda}_0$ and $\bar{\Lambda}_+$.

The orthogonal triangular-block parameterization guarantees that the block restrictions are satisfied, but not the exogeneity restrictions. The second step of the approach thus consists in imposing linear restrictions on the columns of Q_1 to satisfy the exogeneity conditions. To see this, note that from (4.14.45), the exogeneity restrictions (4.14.46) can be expressed as:

$$J \, \bar{H}_0^{-1} \, \bar{e}_j = 0_{h \times 1} \qquad j = 1, \cdots, n-h \tag{4.14.51}$$

with $J = \begin{pmatrix} 0_{h \times n} & I_h \end{pmatrix}$ and $\bar{e}_j$ a $\bar{n}$-dimensional selection vector of zeros that takes a value of 1 on its j^{th} entry. Also, from (4.14.48), (4.14.51) rewrites:

$$J \, \bar{\Lambda}_0^{-1} \, Q' \, \bar{e}_j = 0_{h \times 1} \qquad j = 1, \cdots, n-h \tag{4.14.52}$$

Finally, define $L = \begin{pmatrix} I_n & 0_{n \times h} \end{pmatrix}$. It is easily verified that $Q' \, \bar{e}_j = L' \, Q_1' \, e_j$, with e_j a n-dimensional selection vector of zeros that takes a value of 1 on its j^{th} entry. Then (4.14.52) eventually rewrites:

$$J \, \bar{\Lambda}_0^{-1} \, L' \, Q_1' \, e_j = 0_{h \times 1} \quad \Rightarrow \quad G \, Q_1' \, e_j = 0_{h \times 1} \qquad G \equiv J \, \bar{\Lambda}_0^{-1} \, L' \qquad j = 1, \cdots, n-h \tag{4.14.53}$$

Equation (4.14.53) shows that the exogeneity restrictions are equivalent to linear restrictions on the columns of Q_1. We denote by z_j the number of restrictions on the j^{th} column of Q_1, which is $z_j = h$ for $j = 1, \cdots, n-h$, and $z_j = 0$ for $j = n-h+1, \cdots, n$. Arias et al. (2021) then propose the following algorithm to draw matrices Q_1 and Q_2 that satisfy the exogeneity restrictions:

algorithm 14.10: Construction of a matrix Q satisfying the exogeneity restrictions
For a given matrix $G = J\,\bar{\Lambda}_0^{-1}\,L'$:

1. for $j = 1, \cdots, n$, draw a vector $x_{1,j}$ of dimension $n+1-j-z_j$ from a standard normal distribution and set $w_{1,j} = x_{1,j}/\|x_{1,j}\|$.

2. define $Q_1 = [q_{1,1} \quad \cdots \quad q_{1,n}]$ recursively by $q_{1,j} = K_{1,j} w_{1,j}$, for any matrix $K_{1,j}$ whose columns form an orthonormal basis for the nullspace of the $(j-1+z_j) \times n$ matrix:
   $M_{1,j} = [q_{1,1} \quad \cdots \quad q_{1,j-1} \quad G']'$ for $j = 1, \cdots, n-h$.
   $M_{1,j} = [q_{1,1} \quad \cdots \quad q_{1,j-1}]'$ for $j = n-h+1, \cdots, n$.

3. obtain Q_2 from algorithm 14.4.

4. set $Q = diag(Q_1, Q_2)'$.

Algorithm 14.10 ensures that Q satisfies the exogeneity conditions, but we want to make sure that the relevance conditions are satisfied as well. This implies that the matrix V in (4.14.46) is non-singular. In practice, we may want to make sure that V is, in fact, far from being singular. To do so, Arias et al. (2021) suggest to use a relevance matrix P, defined as:

$$P = (\Gamma_{0,2}^{-1}\Gamma_{0,2}^{-1\prime} + VV')^{-1}VV' \tag{4.14.54}$$

One then checks whether the minimum eigenvalue of P is larger than some chosen λ, with $0 \leq \lambda \leq 1$. This implies that at least λ percent of the variance of any linear combination of the proxys is related to the underlying shocks of interest.

With these elements, it is finally possible to propose a complete Gibbs sampling algorithm for the proxy SVAR:

algorithm 14.11: Gibbs sampling algorithm for the proxy-SVAR

1. set the posterior parameters $\bar{\alpha}$, $\bar{W}$, $\bar{B}$, $\bar{S}$, and H_j, P_j, Q_j for $j = 1, \cdots, \bar{n}$.

2. set the initial value $\bar{\Lambda}_0^{(0)} = I_{\bar{n}}$.

3. at iteration j, draw $\bar{\Lambda}_0^{(j)}$ and $\bar{\Lambda}_+^{(j)}$, using algorithm 14.9.

4. at iteration j, draw $Q^{(j)}$, using algorithm 14.10.

5. obtain the SVAR parameters from $\bar{H}_0^{(j)} = Q^{(j)}\,\bar{\Lambda}_0^{(j)}$ and $\bar{H}_+^{(j)} = Q^{(j)}\,\bar{\Lambda}_+^{(j)}$.

6. given $\bar{H}_0^{(j)}$, compute V from (4.14.46) and the relevance matrix P from (4.14.54); if the minimum eigenvalue of P is larger than λ, keep the draws; else, discard and return to step 3.

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

This concludes the main presentation of the proxy-SVAR methodology. A few additional points are worth noting. First, Arias et al. (2021) argue that as it is, algorithm 14.11 is not sufficient to properly identify the structural shocks of the model. Additional restrictions are required to solve this identification problem, and these can be any among sign, zero, or narrative restrictions.

The case of additional zero restrictions must be handled with care since the exogeneity conditions already involve zero restrictions on the first $n-h$ shocks. Following, a maximum of $n-h-j$ additional zero restrictions can be applied on structural shock j to keep Q well identified. An additional zero restriction on shock j can be expressed as:

$$e_j\, f(\tilde{\Psi})\, L'\, Q_1'\, s_j = G_j\, Q_1'\, s_j = 0 \qquad G_j \equiv e_j\, f(\tilde{\Psi})\, L' \tag{4.14.55}$$

where similarly to section 14.3, the matrix $f(\tilde{\Psi}) = f(\tilde{\Phi})\, h(\Sigma)$ stacks the IRFs for the periods on which the restrictions apply, e_j and s_j are selection vectors with a single 1 entry, and L is defined as in (4.14.53). Stacking the h_j restriction vectors e_j for shock j in a matrix Z_j, this rewrites:

$$Z_j\, f(\tilde{\Psi})\, L'\, Q_1'\, s_j = G_j\, Q_1'\, s_j = 0 \qquad G_j \equiv Z_j\, f(\tilde{\Psi})\, L' \tag{4.14.56}$$

Algorithm 14.10 then needs to be rewritten as follows to account for the additional zero restrictions:

algorithm 14.12: Construction of a matrix Q satisfying the exogeneity and zero restrictions
For given matrices $G = J\, \bar{\Lambda}_0^{-1}\, L'$ and $G_j = Z_j\, f(\tilde{\Psi})\, L'$:

1. for $j = 1, \cdots, n$, draw a vector $x_{1,j}$ of dimension $n+1-j-z_j-h_j$ from a standard normal distribution and set $w_{1,j} = x_{1,j}/\|x_{1,j}\|$.

2. define $Q_1 = [q_{1,1} \quad \cdots \quad q_{1,n}]$ recursively by $q_{1,j} = K_{1,j} w_{1,j}$, for any matrix $K_{1,j}$ whose columns form an orthonormal basis for the nullspace of the $(j-1+z_j+h_j) \times n$ matrix:
   $M_{1,j} = [q_{1,1} \quad \cdots \quad q_{1,j-1} \quad G' \quad G_j']'$ for $j = 1, \cdots, n-h$.
   $M_{1,j} = [q_{1,1} \quad \cdots \quad q_{1,j-1} \quad G_j']'$ for $j = n-h+1, \cdots, n$.

3. obtain Q_2 from algorithm 14.4.

4. set $Q = diag(Q_1, Q_2)'$.

Regarding sign restrictions, Arias et al. (2021) propose a new type of restrictions that apply specifically to the proxy SVAR: **covariance restrictions**. Indeed, the matrix V defined in (4.14.46) represents the covariance matrix between the proxies and the structural shocks to which they relate. Setting restrictions on the signs of the covariances then ensures that only meaningful models will be retained by the Gibbs sampler.

To illustrate this, consider again the case of a proxy SVAR with two proxys correlated with the last two structural shocks of the model. For the first restriction, we want the covariance of the first proxy with the first structural shock (among the last two) to be positive, that is, $V_{11} > 0$. The restriction will hold if:

$$e_1 \times V \times s_1 > 0 \quad \text{or} \quad \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} > 0 \quad \text{or} \quad V_{11} > 0 \tag{4.14.57}$$

The second restriction considers that the covariance between the second proxy and the second structural shock is stronger than that with the first structural shock, that is, $V_{22} > V_{21}$. The restriction will hold if:

$$e_1 \times V \times s_1 > 0 \quad \text{or} \quad \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} > 0 \quad \text{or} \quad V_{22} > V_{21} \tag{4.14.58}$$

Equiped with these additional restrictions, it is possible to define the general Gibbs algorithm for the proxy SVAR:

algorithm 14.13: Gibbs sampling algorithm for the proxy-SVAR with sign and zero restrictions

1. set the posterior parameters $\bar{\alpha}$, $\bar{W}$, $\bar{B}$, $\bar{S}$, and H_j, P_j, Q_j for $j = 1, \cdots, \bar{n}$.
2. set the initial value $\bar{\Lambda}_0^{(0)} = I_{\bar{n}}$.
3. at iteration j, draw $\bar{\Lambda}_0^{(j)}$ and $\bar{\Lambda}_+^{(j)}$, using algorithm 14.9.
4. at iteration j, draw $Q^{(j)}$; use algorithm 14.10 if there are no additional zero restrictions; if additional zero restrictions apply, use algorithm 14.12 instead.
5. obtain the SVAR parameters from $\bar{H}_0^{(j)} = Q^{(j)}\ \bar{\Lambda}_0^{(j)}$ and $\bar{H}_+^{(j)} = Q^{(j)}\ \bar{\Lambda}_+^{(j)}$.
6. given $\bar{H}_0^{(j)}$, compute V from (4.14.46) and the relevance matrix P from (4.14.52); if the minimum eigenvalue of P is larger than λ, continue; else, discard and return to step 3.
7. verify that the sign, narrative and correlation restrictions are satisfied; if yes, keep $\bar{H}_0^{(j)}$ and $\bar{H}_+^{(j)}$; else, discard the draws and return to step 3.
8. repeat until the desired number of iterations is realised.

This concludes the presentation of the proxy SVAR methodology. A final remark applies: similar to the sign restriction methodology, a direct application of algorithm 14.13 does not produce samples from the target distribution (the normal-generalized-normal distribution), and for this reason Arias et al. (2021) propose to apply a similar importance sampling procedure. We do not follow this line for reasons similar to that developed at the end of section 14.3, and apply the simpler and safer algorithm 14.13 without ado.

14.6 How well does the IS-LM model fit postwar E.U. data? (revisited)

This section revisits the E.U. postwar dataset introduced in section 13.6. In the latter, some results appear inconsistent with the stylized predictions of the IS-LM model. One possible reason for this is the simplistic approach used for the exercise: a simple Bayesian VAR with structural identification conducted by Cholesky factorisation.This section introduces a more sophisticated approach: a Bayesian proxy-SVAR with additional sign and covariance restrictions to properly identify the structural shocks.

The base setup is unchanged and includes the data series of real GDP growth, broad money m3, the 3-month interest rate and CPI inflation introduced previously. The dataset is supplemented with two additional series that play the role of proxys for the proxy-SVAR. The first series is a proxy for supply shocks. It is calculated as the quarterly growth rate of the commodity price index supplied by the European Central Bank for the Euro area. The second series is a real demand proxy, obtained from the quarterly growth rate of the OECD Consumer Opinion Surveys index. The correlation between the two series is less than 0.03, making them effectively orthogonal, as expected. The two proxys are represented in Figure 14.1:

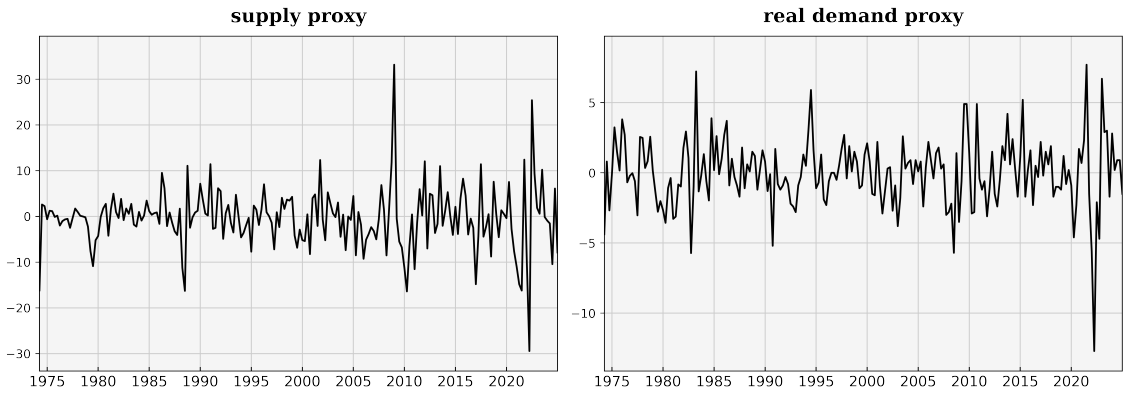


Figure 14.1: Supply and real demand proxys

The new dataset is then used in the proxy SVAR model developed in section 14.5. The relevance parameter λ is set to 0.1 to ensure consistency between the proxys and the identified shocks. Note again that estimating a proxy SVAR is not sufficient in itself to properly identify the structural shocks of the model. To do so additional restrictions are necessary. Table 14.1 summarizes the set of restrictions implemented on impulse response functions:

	supply	money supply	money demand	real demand
gdp	+	+		+
m3				
rate		-	+	
cpi	-	+	+	+

Table 14.1: Sign restrictions on impulse response functions

The restrictions on CPI inflation identify the supply shock, stating it is the only shock that affects inflation downwards. By contrast, all the other shocks are assumed to increase the price level, consistent with traditional Keynesian views. Money supply and money demand shocks are further identified by constraining the latter to result in a rise of the interest rate, while the former contributes to reduce it. Positive restrictions on GDP are set to secure the positive impact of supply, money supply and real demand shocks on short-term economic activity. All the restrictions are set for the initial period of impulse response functions.

It may seem that this setup does not identify the real demand shock, but it is not so. By construction, the real demand shock is correlated with its proxy while the money supply and money demand shocks are orthogonal to it. As the supply shock is identified on its own, this is sufficient for proper identification. Also, to guarantee consistent identification of the structural shocks, positive covariance restrictions are implemented between the supply and real demand shocks and their respective proxys. This minimal setup permits a proper identification of the shocks while leaving a substantial amount of flexibility to the model.

Figure 14.2 reports the impulse response function of the estimated prox-SVAR. Unlike those previously obtained in section 13.6, these responses are consistent with the stylized predictions of the IS-LM model, though by construction for some of them.

All the positive shocks now trigger an increase in real GDP growth. Real demand and monetary shocks are short-lived (about 12 quarters) while supply shocks are significantly longer-lasting (about 24 quarters). This is consistent with the Keynesian view of a permanent effect of supply shocks on production see e.g. Blanchard and Quah (1989)).

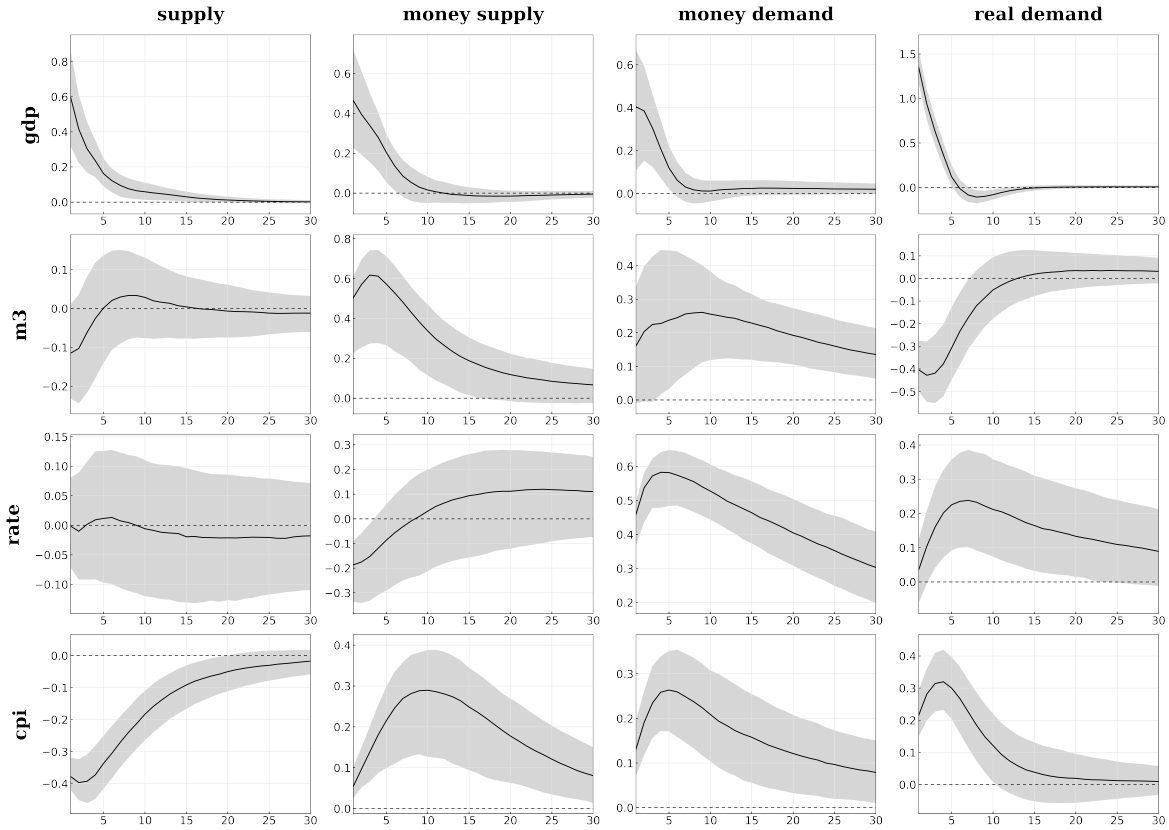


Figure 14.2: Structural impulse response function

Money supply shocks create a transitory drop in the short rate, before it rises again due to improved economic activity. Similarly, supply shocks result in a fall in the 3-month rate, though the effect looks small and non-significant. As expected, the effect is reversed for money demand and real demand shocks. This contrasts with section 13.6, where all the structural shocks were generating rising interest rates.

Broad money m3 increases after a positive money shock, be it supply or demand, the explanation for the latter probably being partial accomodation by the monetary authorities. Supply and real demand shocks both trigger a temporary drop in aggregate money, with no obvious rationale for the latter though the finding is consistent with that of Gali (1992).

Finally, CPI inflation follows common IS-LM wisdom that all shocks result in an increase of the price level, save for supply shocks that result in lower inflation. These results are by construction for the initial period, but not the observed subsequent hump-shaped reaction which reflects the transient effect of structural shocks on the price level.

Figure 14.3 reports the forecast error decomposition for the model. A striking difference with section 13.6 is that at business cycle horizons, fluctuations in real GDP growth are now mostly determined by real demand shocks. Supply shocks now represent the smallest share with barely 10% of the fluctuations, while moneatry shocks roughly account for a quarter of the observed variation. This is much more consistent with the IS-LM framework where IS shocks play a key role in output stabilization.

As expected, broad money remains largely determined by money supply shocks, though money demand account for about 30% of the fluctuations in the long-run, supporting again the hypothesis of partial accomodation by central authorities.

Consistent also with Keynesian theory, the short-term interest rate is overwhelmingly dominated by money demand shocks, with some space left in the long run to money supply shocks. Supply and real demand shocks play almost no role at any horizon, in agreement with the LM curve construction.

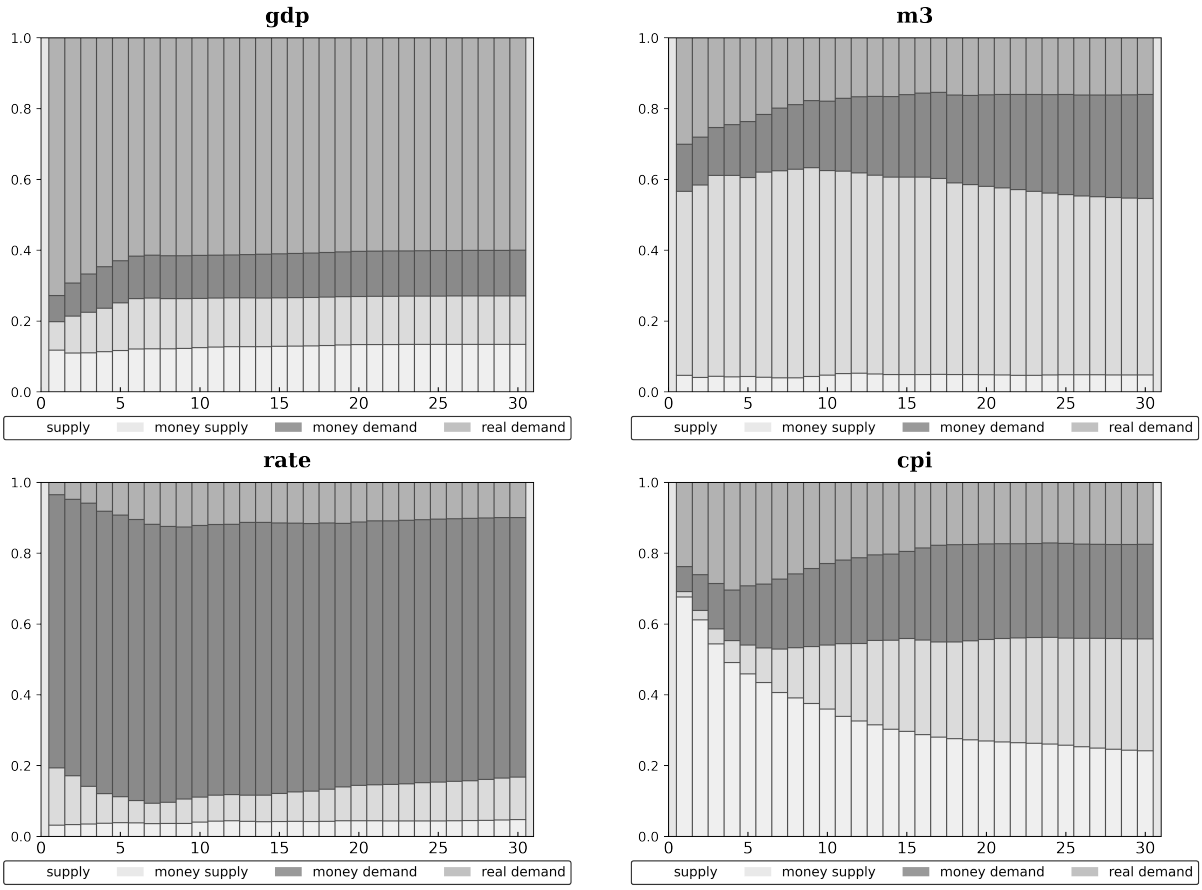


Figure 14.3: Forecast error variance decomposition

Interestingly enough, CPI inflation seems more balanced between the different components. While supply shocks are roughly responsible for 50% of the fluctuations in the short-run, the monetary side gradually takes over. In the long-run the four shocks seem to play more or less at par, implying that fluctuations in the price level may come from multiple sources with varying importance across the sample.

Figure 14.4 displays the historical decomposition for the model. Here again, a striking difference can be observed for the decomposition of GDP fluctuations, compared to section 13.6. The fluctuations now appear to be effectively shared between the different shocks across the sample, with a much larger weight granted to money demand and real demand shocks. This is also true for the recent pandemic crisis, where the decomposition suggests a non-negligible contribution of real demand shocks, both during the crisis and its recovery.

The short-term interest rate remains dominated by the monetary components, the supply side taking the bulk of the fluctuations, and the demand side playing the role of the minority complement. Real demand and supply shocks play almost no role in interest rate determination.

Broad money fluctuations remained almost exclusively dominated by money supply shocks, with a somewhat limited contribution of money demand shocks. Real demand shocks hardly play any role, except during the recent pandemic episode where, interstingly enough, they motivated the initial increase in money mass and also the subsequent cut.

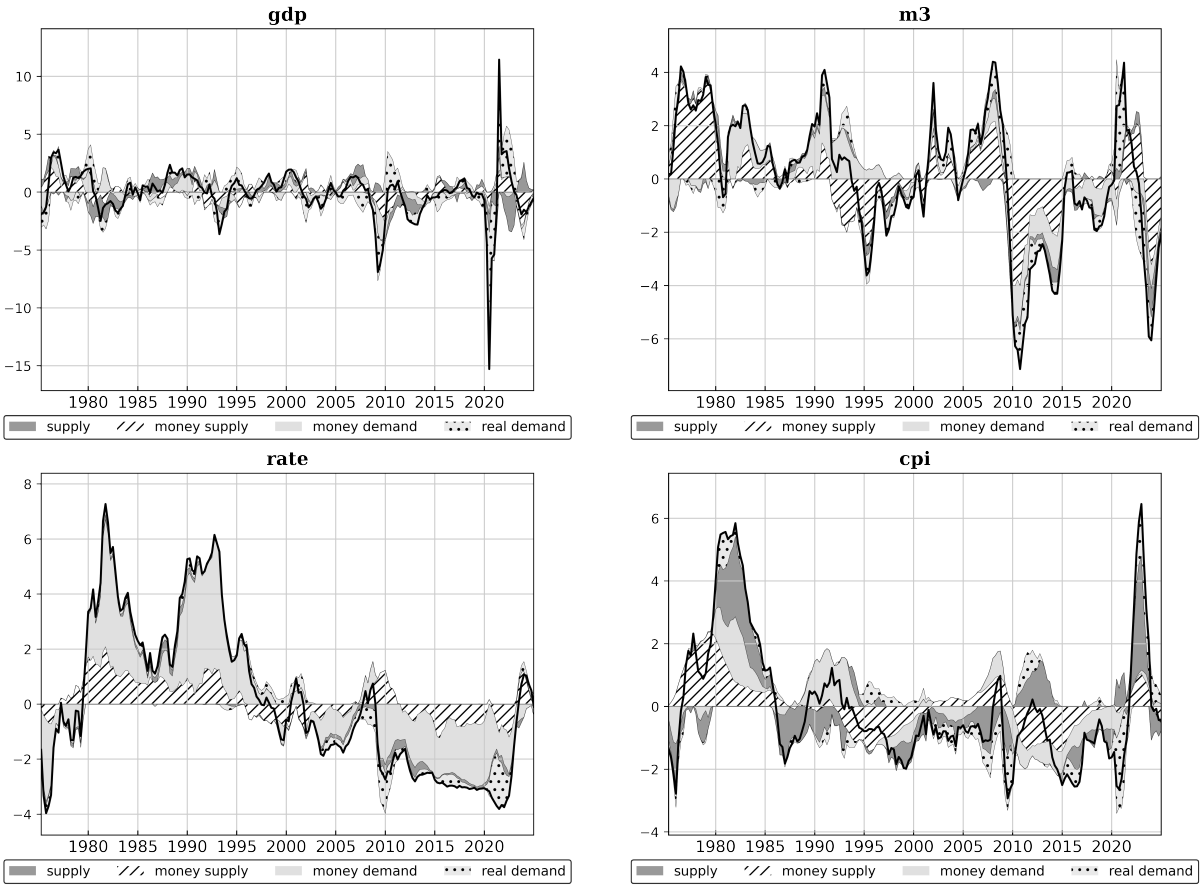


Figure 14.4: Historical decomposition

CPI inflation finally exhibit a much more balanced view than in section 13.6, where supply shocks and money demand shocks played almost no role. Here all the components play a significant role, with varying importance across the sample. For instance, the low inflation of the 2010 decade can be seen as a conjunction of negative money supply contributions (overly restrictive monetary policy) and negative money demand components (insufficient demand for real money balance due to low economic activity). In a very interesting way, the post-pandemic inflation episode now appears as a mix of money supply, real demand, and supply components. This suggests that both the explanations of Bernanke and Blanchard (2023) and Giannone and Primiceri (2023) are correct, but that none of them captures the full picture individually.

Overall, the improved structural identification approach developed in this section makes the model much more consistent with the stylized predictions of the IS-LM framework. Real demand and monetary shocks play a significantly larger role at business cycle horizons, while the influence of the supply side diminishes. Also, all the shocks contribute positively to economic activity, while supply shocks specifically contribute to lower inflation and interest rates at short horizons.

To conclude the exercise, a brief scenario analysis is proposed. As the results obtained so far suggest a strong impact of monetary shocks, one can expect to use monetary policy to enhance economic activity. The exercise thus considers the impact of a substantial cut in the short-term interest rate, droping from 3% at sample end to 2.5% over the next four quarters of the scenario, with an uncertainty of 0.2%. To make sure that the observed cut is the result of monetary policy conducted by central authorities, the scenario uses the structural conditional approach and restricts the conditions to be generated by money supply shocks only.

Figure 14.5 first plots the unconditional forecasts, that is, the forecasts obtained without scenario. The proxy-SVAR model predicts a fall of the interest rate over the next four quarters, but not as steep as the one set by the scenario. The short-term rate can be seen to end the prediction period at about 2.8%, with relatively large credibility bands. To achieve this fall in the 3-month rate, the central authorities are expected to increase the overall money supply over the period, to reach a 5% growth rate at the end of the exercise. The expansionary monetary policy also results in improved economic activity with real GDP growth rising to 1.6% at period end, a modest increase only. CPI inflation remains largely unchanged, with a marginal and non-significant drop over the period stabilizing at 1.9%.

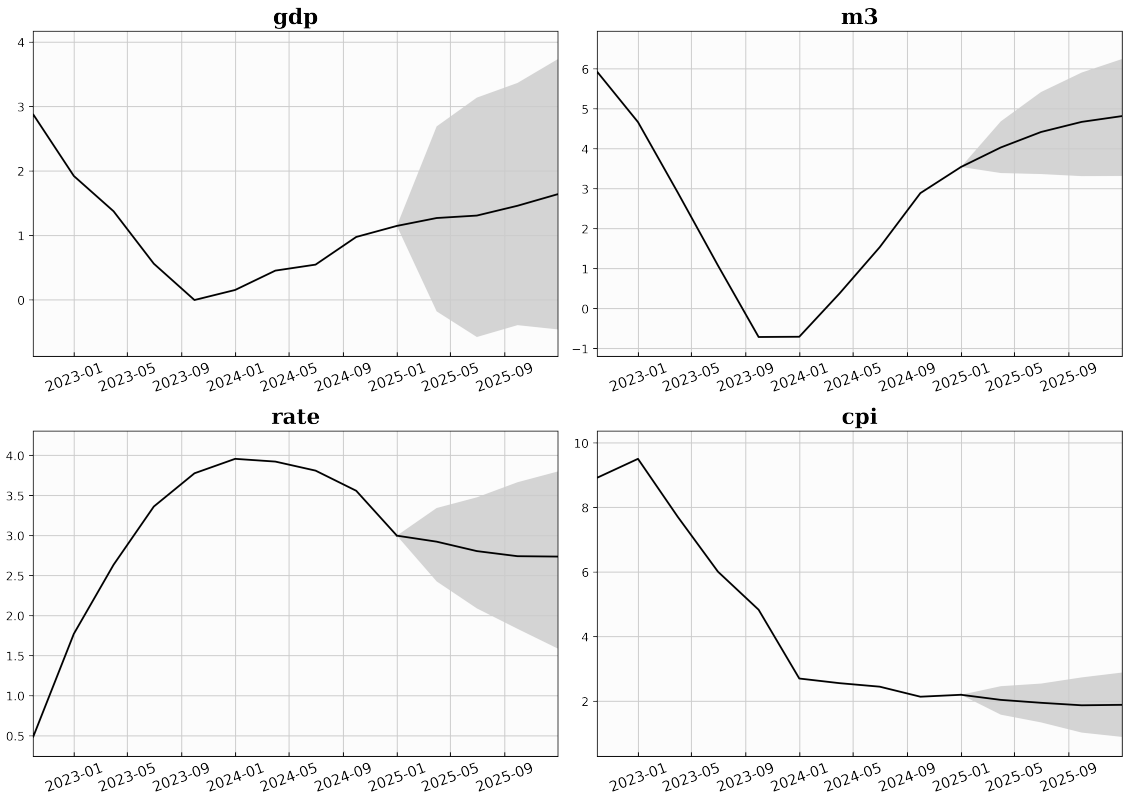


Figure 14.5: Regular forecasts

The scenario is fairly similar to the narrative of the unconditional forecasts, but with a significantly stronger fall in the interest rate. It is thus expected that the qualitative effects will be similar, with magnified quantitative responses. This is indeed what happens. Figure 14.6 reports the conditional forecasts provided by the model.

The interest rate first displays a brutal drop to 2.5%, with 0.2% credibility bands indicating high certainty in the scenario. To sustain this drop the monetary authorities operate a large increase in aggregate money, with m3 growth reaching almost 6% at the end of the period. The strong move in the short rate triggers this time both a large and immediate rise in real GDP growth, escalating then maintaining itself at a 2% rate over the period. This is significantly better than the mild 1.6% observed for the unconditional forecasts. Interestingly enough, this strong expansionary monetary policy does not result in significantly higher inflation, as one might have expected. Inflation remains low at about 2% and plateaus at this value all along the period. Overall, this exercise suggests that there is room for monetary policy as a potent stabilization tool, without an immediate concern on inflationary pressures.

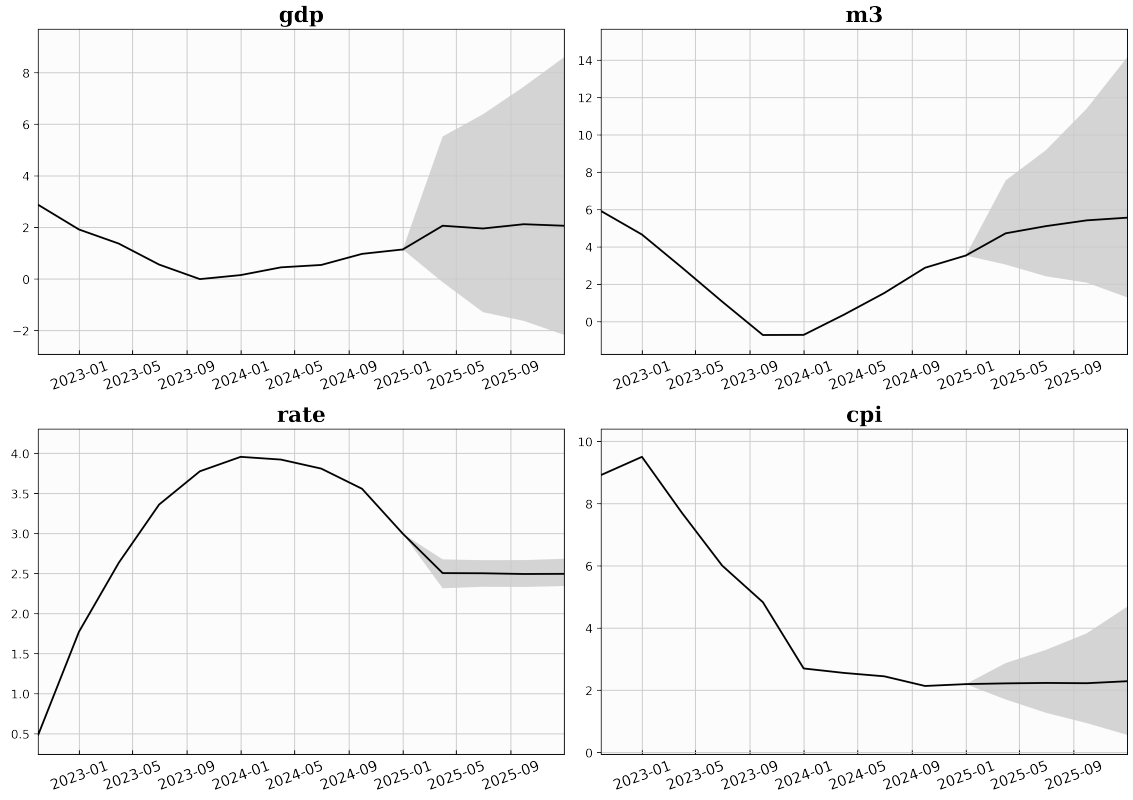


Figure 14.6: Conditional forecasts

Bibliography

Andersson, M., Palmqvist, S., and Waggoner, D. (2010). Density-conditional forecasts in dynamic multivariate models. Working Paper Series 247, Sveriges Riksbank.

Antolin-Diaz, J., Petrella, I., and Rubio-Ramírez, J. F. (2018). Structural scenario analysis with svars. CEPR Discussion Papers 12579, C.E.P.R. Discussion Papers.

Antolin-Diaz, J. and Rubio-Ramírez, J. F. (2018). Narrative sign restrictions for SVARs. *American Economic Review*, 108(10):2802–2829.

Arias, J. E., Rubio-Ramírez, J. F., and Waggoner, D. F. (2018). Inference based on structural vector autoregressions identified with sign and zero restrictions: theory and applications. *Econometrica*, 86(2):685–720.

Arias, J. E., Rubio-Ramírez, J. F., and Waggoner, D. F. (2021). Inference in bayesian proxy-svars. *Journal of Econometrics*, 225(1):88–106.

Banbura, M., Giannone, D., and Lenza, M. (2015). Conditional forecasts and scenario analysis with vector autoregressions for large cross-sections. *International Journal of Forecasting*, 31(3):739–756.

Banbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Bernanke, B. and Blanchard, O. J. (2023). What caused the us pandemic-era inflation? NBER Working papers 31417, National Bureau of Economic Research.

Blanchard, O. J. and Quah, D. (1989). The dynamic effects of aggregate supply and aggregate demand disturbances. *The American Economic Review*, 79(4):655–673.

Caldara, D. and Herbst, E. (2019). Monetary policy, real activity, and credit spreads. *American Economic Journal: Macroeconomics*, 11(1):157–192.

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.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3:1–100.

Fagan, G., Henry, J., and Mestre, R. (2001). An area-wide model (awm) for the euro area. ECB Working papers 42, European Central Bank.

Gali, J. (1992). How well does the is-lm model fit postwar u.s. data? *The Quarterly Journal of Economics*, 107(2):709–738.

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.

Giannone, D., Lenza, M., and Primiceri, G. E. (2019). Priors for the long run. *Journal of the American Statistical Association*, 114(526):565–580.

Giannone, D. and Primiceri, G. (2023). The drivers of post-pandemic inflation. CEPR Working papers 19377, Centre for Economic Policy Research.

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. (2008). *Introduction to Bayesian Econometrics*. Cambridge University Press, 2 edition.

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education, 5 edition.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

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.

Kadiyala, K. R. and Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, 12(2):99–132.

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.

Litterman, R. B. (1985). Forecasting with Bayesian vector autoregressions: five years of experience. Working Papers 274, Federal Reserve Bank of Minneapolis.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer.

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.

Sims, C. A. (1993). A nine variable probabilistic macroeconomic forecasting model. In *Business Cycles, Indicators, and Forecasting*, pages 179–212. University of Chicago Press.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy*, 39:195–214.

Waggoner, D. F. and Zha, T. (1999). Conditional forecasts in dynamic multivariate models. *The Review of Economics and Statistics*, 81(4):639–651.

Waggoner, D. F. and Zha, T. (2003). A gibbs sampler for structural vector autoregressions. *Journal of Economic Dynamics and Control*, 28(2):349–366.

Warne, A., Coenen, G., and Christoffel, K. (2013). Predictive likelihood comparisons with DSGE and DSGE-VAR models. Working Papers 1536, European Central Bank.

Zellner, A. (1996). *An Introduction to Bayesian Inference in Econometrics*. Wiley, 1^{st} edition.

Subject index

Author index