Half Title

Introduction to Bayesian Inference:
A GUIded tour using R

Title Page

Introduction to Bayesian Inference:
A GUIded tour using R

by Andrés Ramírez-Hassan, PhD. Statistical Science.

LOC Page

To my parents, Nancy and Orlando.

Introduction

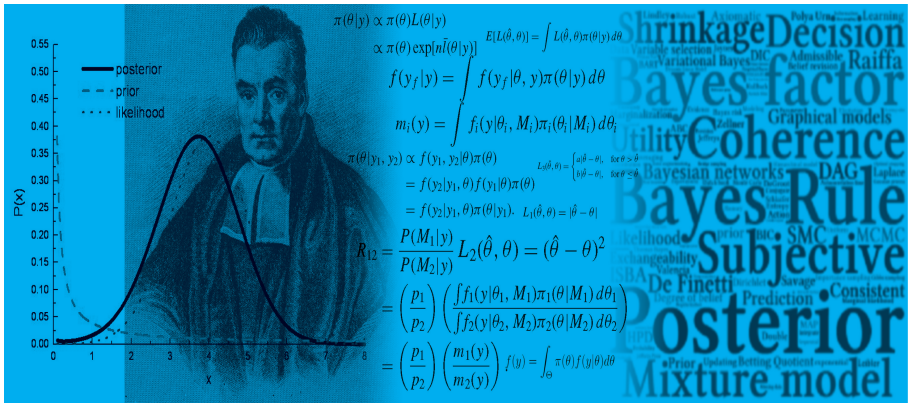

FIGURE 1
Supposedly portrait of Thomas Bayes.

Since late 90's Bayesian inference has gained a lot of popularity among researchers due to the computational revolution and availability of algorithms to solve complex integrals. However, many researchers, students and practitioners still lack understanding and application of this inferential approach. The main reason is the requirement of good programming skills.

Introduction to Bayesian inference: A GUIded tour using R mainly targets those who want to apply Bayesian regression analysis having a good conceptual and formal understanding, but not necessarily having time to develop programming skills. Thus, this book provides a graphical user interface (GUI) to carry out Bayesian regression in a very friendly environment. The book also provides the basic theory, and its code implementation using **R** software [69], some econometric/statistical applications to highlight the potential of Bayesian regression, and theory and computational exercises, for those who are interested in developing more complex models. In particular, this book contains the mathematical proofs step by step of the basic model, which are the base for obtaining the most relevant mathematical results of more complex models.

Our GUI is based on an interactive web application using shiny [16], and some packages in **R**. Users can estimate univariate, multivariate, time series, panel data (longitudinal) and Bayesian model average models using our GUI.

In addition, it gives basic summaries and formal and graphical diagnostics of the posterior chains. Our GUI can be run in any operating system, and is freely available at **https://github.com/besmarter/BSTApp**.

Users can get simulated and real datasets in the folders **DataSim**, and **DataApp**, respectively. The former folder also includes the files that were used to simulate different processes, so, the population parameters are available, and as a consequence, these files can be used as a pedagogical tool to show some statistical properties. The latter folder contains the datasets used in our applications. Users should use these datasets as templates to structuring their own datasets. Simply type **shiny::runGitHub("besmarter/BSTApp" , launch.browser=T)** in the **R** package console or any **R** code editor to run our GUI.[1]

This book has three parts. The first one is about theory (conceptual and mathematical), programming and simulation foundations (chapters 1 to 5), the second part is about applications of regression analysis (chapters 6 to 11), and the third part is about *advanced methods* in Bayesian inference (chapters 12 to 15). I show in some detail the mathematical deductions in the first part of the book, whereas I do not show any proof in the second and third parts. However, same mathematical steps can be used to find the results of parts two and three of the book. I also show three levels regarding computational implementation in the second part of the book: programming ourselves the algorithms, using Bayesian **R** packages, and using our GUI.

Chapter 1 begins with an introduction to formal concepts in Bayesian inference starting with the Bayes' rule, all its components with their formal definitions and basic examples. Then, it presents the basics of Bayesian inference based on decision theory under uncertainty. Chapter 2 presents conceptual differences between Bayesian and Frequentist statistical approaches, and a historical and philosophical perspective about Bayesian statistics and econometrics highlighting differences compared to the Frequentist approach. Chapter 3 presents the differences between the objective and subjective schools in Bayesian inference. Particular attention is put to elicitation techniques, that is, how to transform expert knowledge into prior probabilistic statements. In Chapter 4 I introduce conjugate families in basic statistical models, solving them analytically and computationally. Simulation based methods are shown in Chapter 5, these algorithms are very important in modern Bayesian inference as most realistic models do not have standard forms or analytical solutions. I present our graphical user interface in Chapter 6, and univariate and multivariate regression models are presented in chapters 7 and 8, respectively. Chapter 9 presents the state-space representation of time series models, and Chapter 10 presents Bayesian panel data (longitudinal) models. Chapter 11 introduces Bayesian model averaging. In the third part, there are Chapter 12 introducing hierarchical models, Chapter 13 shows causal inference, Chapter 14 shows Bayesian methods in machine learning algorithms, and

[1] I strongly recommend to type the code line rather than copy and paste it.

Chapter 15 describes some recent methodological developments such as approximate Bayesian computation (ABC), variational Bayes (VB), integrated nested Laplace approximations (INLA), and Bayesian exponential tilted empirical likelihood (BETEL).

About me

My name is Andrés Ramírez-Hassan, I am an applied and theory econometrician working as a Distinguished Professor in the School of Finance, Economics and Government at Universidad EAFIT (Medellín, Colombia). I got a PhD in Statistical Science, a masters degree in Finance, and another in Economics, and also a bachelor's degree in Economics. I was a research fellow at the Department of Econometrics and Business Statistics at Monash University, and a visiting Professor in the Department of Economics at the University of Melbourne and the University of Glasgow. Having completed my PhD degree, much of my research has been in the area of Bayesian Econometrics with applications in crime, finance, health, sports and utilities. My work has been published (or is forthcoming) in the *International Journal of Forecasting*, *Journal of Applied Econometrics*, *Econometric Reviews*, *Journal of Computational and Graphical Statistics*, *The R Journal*, *Economic Modelling*, *Spatial Economic Analysis*, *Economic Inquiry*, *World Development*, *Journal of Sport Economics*, *Empirical Economics*, *Australian and New Zealand Journal of Statistics*, *Brazilian Journal of Probability and Statistics*, and other highly regarded international research outlets.

I founded **BEsmarter** –**B**ayesian **E**conometrics: **s**imulations, **m**odels and **a**pplications to **research**, **teaching** and **e**ncoding with **r**esponsibility–. This is a research group whose **mission** is to *lead and excel in the generation and dissemination of Bayesian Econometric knowledge through research, teaching and software*. We en**vision** *worldwide econometric research, teaching and applications based on the Bayesian framework that*:

- Inspires new econometric ideas
- Creates a user friendly environment for applications of Bayesian econometrics
- Transforms classic econometric research, teaching and applications
- And where one of the main concerns of science is to solve social problems

mail: aramir21@gmail.com / aramir21@eafit.edu.co

website: **http://www.besmarter-team.org** / **https://sites.google.com/view/arh-bayesian**

FIGURE 2

Contents

Foreword

Preface

The main goal of this book is to make more approachable the Bayesian inferential framework to students, researchers and practitioners who want to understand and apply this statistical/econometric approach, but who do not have time to develop programming skills. I tried to have a balance between applicability and theory. Then, this book comes with a very friendly graphical user interface (GUI) to implement the most common regression models, but also contains the basic mathematical developments, as well as their code implementation, for those who are interested in advancing in more complex models.

To instructors and students

This book is divided in three parts, foundations (chapters 1 to 5), regression analysis (chapters 6 to 11), and *Advanced* methods (chapters 12 to 15). Our graphical user interface (GUI) targets the second part. This can be download at **https://github.com/besmarter/BSTApp**. Instructors and students can have all codes, simulated and real data sets are also there. To install our GUI just type **shiny::runGitHub("besmarter/BSTApp" , launch.browser=T)** in the **R** package console or any **R** code editor, and execute it.

Students should have some basic knowledge in probability theory and statistics, particularly, regression analysis. It is strongly recommended to have some familiarity with standard univariate and multivariate probability distributions.

I included some formal and computational exercises at the end of each chapter. This would help students to have a better understanding of the material shown in each chapter. A manual with the solutions of exercises accompanies this book.

Instructors can use this book as a text in a course of introduction to Bayesian Econometrics/Statistics with a high emphasis on implementation and applications. This book is complementary, rather than substitute, of excellent books in the topic such as [32, 77, 40, 37, 51] and [50].

Acknowledgments

I started our GUI in the 2016 after being diagnosed with cervical dystonia. I used to work in this side project on weekends, I named this time "nerd weekends", and it was a kind of release from my health condition. Once I got better, I invited Mateo Graciano, my former student, business partner and friend, to be part of the project, he helped me a lot developing our GUI, and I am enormously thankful to Mateo. I would also like to thank members

of the BEsmarter research group from Universidad EAFIT, and NUMBATs members from Monash University for your comments and recommendations to improve our GUI.

This book is an extension of the paper *A GUIded tour of Bayesian regression* [75], which is a brief user guide of our GUI. So, I decided to write this book to show the underlying theory and codes in our GUI, and use it as a text book in my course in Bayesian econometrics/statistics. I acknowledge and offer my gratitude to my students in this subject, their insight and thoughtful questions have helped me to get a better understanding of this material.

I also thank Chris Parmeter for your suggestions about how to present our user guide, Professor Raul Pericchi and Juan Carlos Correa who introduced me to Bayesian statistics, Liana Jacobi and Chun Fung Kwok (Jackson) from the University of Melbourne and David Frazier from Monash University for nice talks and amazing collaborations in Bayesian Econometrics/statistics, Professor Peter Diggle to support my career, and particularly, Professor Gael Martin, who gave me a chance to work with her, she is an inspiring intellectual figure. Finally, my colleagues and staff from Universidad EAFIT have always given me their support.

To my parents, Orlando and Nancy, who have given me their unconditional support. They have taught me that the primary aspect of the human being's spiritual evolution is humility. I am in my way to learn this.

Symbols

Symbol Description

$\neg$	Negation symbol	argmax	Argument of the maximum
$\propto$	Proportional symbol	argmin	Argument of the minimum
$\perp$	Independence symbol	tr	Trace operator
$\mathcal{R}$	The Real set	vec	Vectorization operator
$\emptyset$	Empty set	lim	Limit
$\mathbb{1}$	Indicator function	$\otimes$	Kronecker product
P	Probability measure	diag$\{\cdot\}$	Diagonal matrix
$:=$	Is defined as	$dim\{\cdot\}$	Dimension of an object

Part I

Foundations: Theory, simulation methods and programming

1

Basic formal concepts

We introduce formal concepts in Bayesian inference starting with the Bayes' rule, all its components with their formal definitions and basic examples. In addition, we present some nice features of Bayesian inference such as Bayesian updating, and asymptotic sampling properties, and the basics of Bayesian inference based on decision theory under uncertainty, presenting important concepts like loss function, risk function and optimal rules.

1.1 The Bayes' rule

As expected the point of departure to perform Bayesian inference is the Bayes' rule,[1] which is the Bayes' solution to the inverse probability of causes, this rule combines prior beliefs with objective probabilities based on repeatable experiments. In this way, we can move from observations to probable causes.

Formally, the conditional probability of A_i given B is equal to the conditional probability of B given A_i times the marginal probability of A_i over the marginal probability of B,

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i, B)}{P(B)} \\ &= \frac{P(B|A_i) \times P(A_i)}{P(B)}, \end{aligned} \tag{1.1}$$

where by the law of total probability $P(B) = \sum_i P(B|A_i)P(A_i) \neq 0$, $\{A_i, i = 1, 2, \dots\}$ is a finite or countably infinite partition of a sample space.

In the Bayesian framework, B is sample information that updates a probabilistic statement about an unknown object A_i following probability rules. This is done by means of the Bayes' rule using prior "beliefs" about A_i, that is, $P(A_i)$, sample information relating B to the particular state of the nature A_i through a probabilistic statement, $P(B|A_i)$, and the probability of observing that specific sample information $P(B)$.

[1] Observe that I use the term "Bayes' rule" rather than "Bayes' theorem". It was Laplace [53] who actually generalized the Bayes' theorem [6]. His generalization is named the Bayes' rule.

Let's see a simple example, *the base rate fallacy*:

Assume that the sample information comes from a positive result from a test whose true positive rate (sensitivity) is 98%, $P(+|\text{disease}) = 0.98$. On the other hand, the prior information regarding being infected with this disease comes from a base incidence rate that is equal to 0.002, that is $P(\text{disease}) = 0.002$. Then, *what is the probability of being actually infected?*

This is an example of *the base rate fallacy*, where having a positive test result from a disease whose base incidence rate is tiny gives a low probability of actually having the disease.

The key to answer the question is based on understanding the difference between the probability of having the disease given a positive result, $P(\text{disease}|+)$, versus the probability of a positive result given the disease, $P(+|\text{disease})$. The former is the important result, and the Bayes' rule help us to get the answer. Using the Bayes' rule (equation 1.1):

$$\begin{aligned} P(\text{disease}|+) &= \frac{P(+|\text{disease}) \times P(\text{disease})}{P(+)} \\ &= \frac{0.98 \times 0.002}{0.98 \times 0.002 + (1 - 0.98) \times (1 - 0.002)} \\ &= 0.09, \end{aligned}$$

where $P(+) = P(+|\text{disease}) \times P(\text{disease}) + P(+|\neg\text{disease}) \times P(\neg\text{disease})$.[2]

R code. The base rate fallacy

```
PD <- 0.002 # Probability of disease
PPD <- 0.98 # True positive (Sensitivity)
PDP <- PD * PPD / (PD * PPD + (1 - PD)*(1 - PPD))
paste("Probability of disease given a positive test is", sep
      = " ", round(PDP, 2))
"Probability of disease given a positive test is 0.09"
```

We observe that despite of having a positive result, the probability of having the disease is low. This due to the base rate being tiny.

Another interesting example, which is at the heart of the origin of the Bayes' theorem [6], is related to the existence of God [86]. The Section X of David Hume's "An Inquiry concerning Human Understanding, 1748" is named *Of Miracles*. There, Hume argues that when someone claims to have seen a

[2] $\neg$ is the negation symbol. In addition, we have that $P(B|A) = 1 - P(B|A^c)$ in this example, where A^c is the complement of A. However, it is not always the case that $P(B|A) \neq 1 - P(B|A^c)$.

miracle, this is poor evidence it actually happened, since it goes against what we see every day. Then, Richard Price, who actually finished and published "An essay towards solving a problem in the doctrine of chances" in 1763 after Bayes died in 1761, argues against Hume saying that there is a huge difference between *impossibility* as used commonly in conversation and *physical impossibility*. Price used an example of a dice with a million sides, where *impossibility* is getting a particular side when throwing this dice, and *physical impossibility* is getting a side that does not exist. In millions throws, the latter case never would occur, but the former eventually would.

Let's say that there are two cases of resurrection (Res), Jesus Christ and Elvis, and the total number of people who have ever lived is 108.5 billion,[3] then the prior base rate is $2/(108.5 \times 10^9)$. On the other hand, let's say that the sample information comes from a very reliable witness whose true positive rate is 0.9999999. Then, *what is the probability of this miracle?*[4]

Using the Bayes' rule:

$$\begin{aligned} P(\text{Res}|\text{Witness}) &= \frac{P(\text{Witness}|\text{Res}) \times P(\text{Res})}{P(\text{Witness})} \\ &= \frac{2/(108.5 * 10^9) \times 0.9999999}{2/(108.5 * 10^9) \times 0.9999999 + (1 - 2/(108.5 * 10^9)) \times (1 - 0.9999999)} \\ &= 0.000184297806959661 \end{aligned}$$

where $P(\text{Witness}) = P(\text{Witness}|\text{Res}) \times P(\text{Res}) + (1 - P(\text{Witness}|\text{Res})) \times (1 - P(\text{Res}))$.

Thus, 1.843×10^{-4} is the probability of a resurrection given a very reliable witness.

R code. Of Miracles

```
# Probability of resurrection
PR <- 2/(108.5 * 10^9)
PWR <- 0.9999999 # True positive rate
PRW <- PR * PWR / (PR * PWR + (1 - PR)*(1 - PWR))
paste("Probability of resurrection given witness is", sep =
    " ", PRW)
"Probability of resurrection given witness is
    0.000184297806959661"
```

Observe that we can condition on many events in the Bayes' rule. Let's have two conditioning events B and C, then equation 1.1 becomes

[3] https://www.wolframalpha.com/input/?i=number+of+people+who+have+ever+lived+on+Earth

[4] https://www.r-bloggers.com/2019/04/base-rate-fallacy-or-why-no-one-is-justified-to-believe-that-jesus-rose/

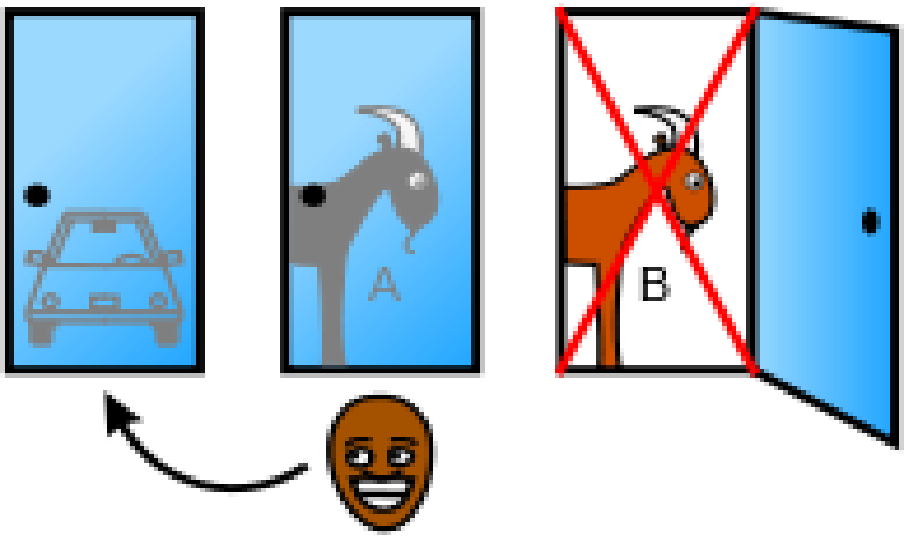

FIGURE 1.1
The Monty Hall problem.

$$P(A_i|B,C) = \frac{P(A_i,B,C)}{P(B,C)} = \frac{P(B|A_i,C) \times P(A_i|C) \times P(C)}{P(B|C)P(C)}. \quad (1.2)$$

Let's use this rule in one of the most intriguing statistical puzzles, *the Monty Hall problem*, to illustrate how to use equation 1.2 [81, 82]. This was the situation faced by a contestant in the American television game show *Let's Make a Deal*. There, the contestant was asked to choose a door where behind one door there is a car, and behind the others, goats. Let's say that the contestant picks door No. 1, and the host (Monty Hall), who knows what is behind each door, opens door No. 3, where there is a goat (see Figure 1.1). Then, the host asks the tricky question to the contestant, *do you want to pick door No. 2?*

Let's name P_i the event **contestant picks door No.** i, which stays close, H_i the event **host picks door No.** i, which is open, and there is a goat, and C_i the event **car is behind door No.** i. In this particular setting, the contestant is interested in the probability of the event $P(C_2|H_3, P_1)$. A naive answer would be that it is irrelevant as initially $P(C_i) = 1/3,\ i = 1, 2, 3$, and now $P(C_i|H_3) = 1/2,\ i = 1, 2$ as the host opened door No. 3. So, why bothering changing the initial guess if the odds are the same (1:1)? The important point here is that the host knows what is behind each door, and picks a door where there is a goat given contestant choice. In this particular

setting, $P(H_3|C_3, P_1) = 0$, $P(H_3|C_2, P_1) = 1$ and $P(H_3|C_1, P_1) = 1/2$. Then, using equation 1.2

$$\begin{aligned} P(C_2|H_3, P_1) &= \frac{P(C_2, H_3, P_1)}{P(H_3, P_1)} \\ &= \frac{P(H_3|C_2, P_1)P(C_2|P_1)P(P_1)}{P(H_3|P_1) \times P(P_1)} \\ &= \frac{P(H_3|C_2, P_1)P(C_2)}{P(H_3|P_1)} \\ &= \frac{1 \times 1/3}{1/2}, \end{aligned}$$

where the third equation uses the fact that C_i and P_i are independent events, and $P(H_3|P_1) = 1/2$ due to this depending just on P_1 (not on C_2).

Therefore, changing the initial decision increases the probability of getting the car from 1/3 to 2/3! Thus, it is always a good idea to change the door.

Let's see a simulation exercise to check this answer:

R code. The Monty Hall problem

```
set.seed(0101) # Set simulation seed
S <- 100000 # Simulations
Game <- function(switch = 0){
  # switch = 0 is not change
  # switch = 1 is to change
  opts <- 1:3
  car <- sample(opts, 1) # car location
  guess1 <- sample(opts, 1) # Initial guess

  if(car != guess1) {
   host <- opts[-c(car, guess1)]
  } else {
   host <- sample(opts[-c(car, guess1)], 1)
  }
  win1 <- guess1 == car # Win no change
  guess2 <- opts[-c(host, guess1)]
  win2 <- guess2 == car # Win change
  if(switch == 0){
    win <- win1
  } else {
    win <- win2
  }
  return(win)
}

#Win probabilities not changing
Prob <- mean(replicate(S, Game(switch = 0)))
Prob
0.3334

#Win probabilities changing
Prob <- mean(replicate(S, Game(switch = 1)))
Prob
0.6654
```

1.2 Bayesian framework: A brief summary of theory

For two random objects $\boldsymbol{\theta}$ and $\mathbf{y}$, the Bayes' rule may be analogously used,[5]

[5]From a Bayesian perspective θ is fixed, but unknown. Then, it is treated as a random object despite the lack of variability (see Chapter 2).

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{1.3}$$

where $\pi(\boldsymbol{\theta}|\mathbf{y})$ is the posterior density function, $\pi(\boldsymbol{\theta})$ is the prior density, $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood (statistical model), and

$$p(\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \mathbb{E}\left[p(\mathbf{y}|\boldsymbol{\theta})\right] \tag{1.4}$$

is the marginal likelihood or prior predictive. Observe that for this expected value to be meaningful the prior should be a proper density, that is, integrates to one, otherwise, it does not make sense.

Observe that $p(\mathbf{y}|\boldsymbol{\theta})$ is not a density in $\boldsymbol{\theta}$. In addition, $\pi(\boldsymbol{\theta})$ does not have to integrate to 1, that is, $\pi(\boldsymbol{\theta})$ can be an improper density function, $\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$. However, $\pi(\boldsymbol{\theta}|\mathbf{y})$ is a proper density function, that is, $\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = 1$. For instance, set $\pi(\boldsymbol{\theta}) = c$, where c is a constant, then $\int_{\boldsymbol{\Theta}} cd\boldsymbol{\theta} = \infty$. However, $\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \frac{p(\mathbf{y}|\boldsymbol{\theta})\times c}{\int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta})\times cd\boldsymbol{\theta}} d\boldsymbol{\theta} = 1$ where c cancels out.

$\pi(\boldsymbol{\theta}|\mathbf{y})$ is a sample updated "probabilistic belief" version of $\pi(\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta})$ is a prior probabilistic belief which can be constructed from previous empirical work, theory foundations, expert knowledge and/or mathematical convenience (see Chapters 3 and 4). This prior usually depends on parameters, which are named *hyperparameters*. In addition, the Bayesian approach implies using a probabilistic model about $\mathbf{y}$ given $\boldsymbol{\theta}$, that is, $p(\mathbf{y}|\boldsymbol{\theta})$, where its integral over $\boldsymbol{\Theta}$, $p(\mathbf{y})$ is named *the model evidence* due to being a measure of model fit to the data.

Observe that the Bayesian inferential approach is conditional, that is, what can we learn about an unknown object $\boldsymbol{\theta}$ given that we already observed $\mathbf{y}$? The answer is also conditional on the probabilistic model, that is $p(\mathbf{y}|\boldsymbol{\theta})$. So, what if we want to compare different models, let's say $\mathcal{M}_m$, $m = \{1, 2, \ldots, M\}$. Then, we should make explicit this in the Bayes' rule formulation,

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_m) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}_m) \times \pi(\boldsymbol{\theta}|\mathcal{M}_m)}{p(\mathbf{y}|\mathcal{M}_m)}. \tag{1.5}$$

The posterior model probability is

$$\pi(\mathcal{M}_m|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{M}_m) \times \pi(\mathcal{M}_m)}{p(\mathbf{y})}, \tag{1.6}$$

where $p(\mathbf{y}|\mathcal{M}_m) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}_m) \times \pi(\boldsymbol{\theta}|\mathcal{M}_m)d\boldsymbol{\theta}$ due to equation 1.5, and $\pi(\mathcal{M}_m)$ is the prior model probability.

Calculating $p(\mathbf{y})$ in equations 1.3 and 1.6 is very demanding most of the realistic cases. Fortunately, it is not required when performing inference about $\boldsymbol{\theta}$ as this is integrated out from it. Then, all what you need to know about the

shape of $\boldsymbol{\theta}$ is in $p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}_m) \times \pi(\boldsymbol{\theta}|\mathcal{M}_m)$ or without explicitly conditioning on $\mathcal{M}_m$,

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}). \tag{1.7}$$

Equation 1.7 is a very good shortcut to perform Bayesian inference about $\boldsymbol{\theta}$.

We also can avoid calculating $p(\mathbf{y})$ when performing model selection (hypothesis testing) using posterior odds ratio, that is, comparing models $\mathcal{M}_1$ and $\mathcal{M}_2$,

$$\begin{aligned} PO_{12} &= \frac{\pi(\mathcal{M}_1|\mathbf{y})}{\pi(\mathcal{M}_2|\mathbf{y})} \\ &= \frac{p(\mathbf{y}|\mathcal{M}_1)}{p(\mathbf{y}|\mathcal{M}_2)} \times \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)}, \end{aligned} \tag{1.8}$$

where the first term in equation 1.8 is named the Bayes factor, and the second term is the prior odds. Observe that the Bayes factor is a ratio of ordinates for $\mathbf{y}$ under different models. Then, the Bayes factor is a measure of relative sample evidence in favor of model 1 compared to model 2.

However, we still need to calculate $p(\mathbf{y}|\mathcal{M}_m) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}_m)\pi(\boldsymbol{\theta}|\mathcal{M}_m)d\boldsymbol{\theta} = \mathbb{E}\left[p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}_m)\right]$. For this integral to be meaningful, the prior must be proper. Using improper prior has unintended consequences when comparing models, for instance, parsimonious models are favored by posterior odds or Bayes factors depend on units of measure (see Chapter 4).

A nice feature of comparing models using posterior odds is that if we have an exhaustive set of competing models such that $\sum_{m=1}^{M} \pi(\mathcal{M}_m|\mathbf{y}) = 1$, then we can recover $\pi(\mathcal{M}_m|\mathbf{y})$ without calculating $p(\mathbf{y})$. In particular, given two models $\mathcal{M}_1$ and $\mathcal{M}_2$ such that $\pi(\mathcal{M}_1|\mathbf{y}) + \pi(\mathcal{M}_2|\mathbf{y}) = 1$. Then, $\pi(\mathcal{M}_1|\mathbf{y}) = \frac{PO_{12}}{1+PO_{12}}$ and $\pi(\mathcal{M}_2|\mathbf{y}) = 1 - \pi(\mathcal{M}_1|\mathbf{y})$. In general, $\pi(\mathcal{M}_m|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{M}_m)\times\pi(\mathcal{M}_m)}{\sum_{l=1}^{M} p(\mathbf{y}|\mathcal{M}_l)\times\pi(\mathcal{M}_l)}$. These posterior model probabilities can be used to perform Bayesian model averaging.

Table 1.1 shows guidelines for the interpretation of $2\log(PO_{12})$ [49]. This transformation is done to replicate the structure of the likelihood ratio test statistic. However, posterior odds do not require nested models as the likelihood ratio test does.

Observe that the posterior odds ratio is a relative criterion, that is, we specify an exhaustive set of competing models, and compare them. However, we may want to check the performance of a model in its own or use a non-informative prior. In this case, we can use *the posterior predictive p-value* [30, 31].[6]

[6] [3] show potential issues due to using data twice in the construction of the predictive p values. They also present alternative proposals, for instance, *the partial posterior predictive p value.*

TABLE 1.1
Kass and Raftery guidelines.

$2 \times \log(PO_{12})$	PO_{12}	**Evidence against $\mathcal{M}_2$**
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

The intuition behind the predictive p-value is simple: analyze discrepancy between model's assumptions and data by checking a potential extreme tail-area probability. Observe that this approach does not check if a model is true, its focus is on potential discrepancies between a model and the data at hand.

This is done simulating pseudo-data from our sampling model ($\mathbf{y}^{(s)}, s = 1, 2, \ldots, S$) using draws from the posterior distribution, and then calculating a discrepancy measure, $D(\mathbf{y}^{(s)}, \boldsymbol{\theta})$, to estimate the posterior predictive p-value, $p_D(\mathbf{y}) = P[D(\mathbf{y}^{(s)}, \boldsymbol{\theta}) \geq D(\mathbf{y}, \boldsymbol{\theta})]$ using the proportion of the S draws for which $D(\mathbf{y}^{(s)}, \boldsymbol{\theta}^{(s)}) \geq D(\mathbf{y}, \boldsymbol{\theta}^{(s)})$. Extreme tail probabilities ($p(D_{\mathbf{y}}) \leq 0.05$ or $p(D_{\mathbf{y}}) \geq 0.95$) suggest potential discrepancy between the data and the model. [31] also suggest the posterior predictive p-value based on the *minimum discrepancy*, $D_{min}(\mathbf{y}) = \min_{\boldsymbol{\theta}} D(\mathbf{y}, \boldsymbol{\theta})$, and the *average discrepancy* statistic $D(\mathbf{y}) = \mathbb{E}[D(\mathbf{y}, \boldsymbol{\theta})] = \int_{\boldsymbol{\Theta}} D(\mathbf{y}, \boldsymbol{\theta}) \pi(\theta|\mathbf{y}) d\boldsymbol{\theta}$. These alternatives can be more computational demanding.

The Bayesian approach is also suitable to get probabilistic predictions, that is, we can obtain a posterior predictive density

$$\begin{aligned} \pi(\mathbf{Y}_0|\mathbf{y}, \mathcal{M}_m) &= \int_{\boldsymbol{\Theta}} \pi(\mathbf{Y}_0, \boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_m) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \pi(\mathbf{Y}_0|\boldsymbol{\theta}, \mathbf{y}, \mathcal{M}_m) \pi(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_m) d\boldsymbol{\theta}. \end{aligned} \tag{1.9}$$

Observe that equation 1.9 is again an expectation $\mathbb{E}[\pi(\mathbf{Y}_0|\boldsymbol{\theta}, \mathbf{y}, \mathcal{M}_m)]$, this time using the posterior distribution. Therefore, the Bayesian approach takes estimation error into account when performing prediction.

As we have shown many times, expectation (integration) is a common feature in Bayesian inference. That is why the remarkable relevance of computation based on *Monte Carlo integration* in the Bayesian framework (see Chapter 5).

Bayesian model average (BMA) allows considering model uncertainty in prediction or any unknown probabilistic object. In the prediction case,

$$\pi(\mathbf{Y}_0|\mathbf{y}) = \sum_{m=1}^{M} \pi(\mathcal{M}_m|\mathbf{y}) \pi(\mathbf{Y}_0|\mathbf{y}, \mathcal{M}_m), \tag{1.10}$$

and parameters case,

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \sum_{m=1}^{M} \pi(\mathcal{M}_m|\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_m), \tag{1.11}$$

where

$$\mathbb{E}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{m=1}^{M} \hat{\boldsymbol{\theta}}_m \pi(\mathcal{M}_m|\mathbf{y}), \tag{1.12}$$

and

$$Var(\boldsymbol{\theta}|\mathbf{y}) = \sum_{m=1}^{M} \pi(\mathcal{M}_m|\mathbf{y})\widehat{Var}(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_m) + \sum_{m=1}^{M} \pi(M_m|\mathbf{y})(\hat{\boldsymbol{\theta}}_m - \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}])^2, \tag{1.13}$$

$\hat{\boldsymbol{\theta}}_m$ and $\widehat{Var}(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_m)$ are the posterior mean and variance under model m, respectively.

Observe how the variance in equation 1.13 encloses extra variability due to potential differences between mean posterior estimates associated with each model, and the posterior mean involving model uncertainty in equation 1.12.

A nice advantage of the Bayesian approach, which is very useful in *state space representations* (see Chapter 9), is the way that the posterior distribution updates with new sample information. Given $\mathbf{y} = \mathbf{y}_{1:t+1}$ a sequence of observations from 1 to $t+1$, then

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{y}_{1:t+1}) &\propto p(\mathbf{y}_{1:t+1}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \\ &= p(y_{t+1}|\mathbf{y}_{1:t}, \boldsymbol{\theta}) \times p(\mathbf{y}_{1:t}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \\ &\propto p(y_{t+1}|\mathbf{y}_{1:t}, \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}|\mathbf{y}_{1:t}). \end{aligned} \tag{1.14}$$

We observe that the new prior is just the posterior distribution using the previous observations. This is particular useful under the assumption of *conditional independence*, that is, $y_{t+1} \perp \mathbf{y}_{1:t}|\boldsymbol{\theta}$, then $p(y_{t+1}|\mathbf{y}_{1:t}, \boldsymbol{\theta}) = p(y_{t+1}|\boldsymbol{\theta})$ such that the posterior can be recovered recursively [67]. This facilities online updating due to all information up to t being in $\boldsymbol{\theta}$. Then, $\pi(\boldsymbol{\theta}|\mathbf{y}_{1:t+1}) \propto p(y_{t+1}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}|\mathbf{y}_{1:t}) \propto \prod_{h=1}^{t+1} p(y_h|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})$. This recursive expression can be calculated faster at some specific point in time t compared to a batch mode algorithm, which requires processing simultaneously all information up to t.

It is also important to wonder about the sampling properties of "Bayesian estimators". This topic has attracted attention of statisticians and econometricians long time ago. For instance, asymptotic posterior concentration at the population parameter vector is discussed by [12]. Convergence of posterior distributions is stated by the Bernstein-von Mises theorem [54], which

creates a link between *credible intervals (sets)* and confidence intervals (sets), where a credible interval is an interval in the domain of the posterior distribution within which an unknown parameter falls with a particular probability. Credible intervals treat bounds as fixed and parameters as random, whereas confidence intervals reverse this. There are many settings in parametric models where Bayesian credible intervals with α level converge asymptotically to confidence intervals at α level. This suggests that Bayesian inference is asymptotically correct from a sampling perspective in these settings.

A heuristic approach to show this in the simplest case where we assume random sampling and $\theta \in \mathcal{R}$ is the following: $p(\mathbf{y}|\theta) = \prod_{i=1}^{N} p(y_i|\theta)$ such that the log likelihood is $l(\mathbf{y}|\theta) \equiv \log p(\mathbf{y}|\theta) = \sum_{i=1}^{N} \log p(y_i|\theta) = N \times \bar{l}(\mathbf{y}|\theta)$ where $\bar{l} \equiv \frac{1}{N} \sum_{i=1}^{N} \log p(y_i|\theta)$ is the mean likelihood.[7] Then, the posterior distribution is proportional to

$$\begin{aligned} \pi(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta) \times \pi(\theta) \\ &= \exp\left\{N \times \bar{l}(\mathbf{y}|\theta)\right\} \times \pi(\theta). \end{aligned} \tag{1.15}$$

Observe that as the sample size gets large, that is, $N \to \infty$, the exponential term should dominate the prior distribution as long as this does not depend on N such that the likelihood determines the posterior distribution asymptotically.

Maximum likelihood theory shows that $\lim_{N\to\infty} \bar{l}(\mathbf{y}|\theta) \to \bar{l}(\mathbf{y}|\theta_0)$ where θ_0 is the population parameter of the data generating process. In addition, doing a second order Taylor expansion of the log likelihood at the Maximum likelihood estimator,

$$\begin{aligned} l(\mathbf{y}|\theta) &\approx l(\mathbf{y}|\hat{\theta}) + \left.\frac{dl(\mathbf{y}|\theta)}{d\theta}\right|_{\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2} \left.\frac{d^2 l(\mathbf{y}|\theta)}{d\theta^2}\right|_{\hat{\theta}} (\theta - \hat{\theta})^2 \\ &= l(\mathbf{y}|\hat{\theta}) + \frac{1}{2} \sum_{i=1}^{N} \left.\frac{d^2 l(y_i|\theta)}{d\theta^2}\right|_{\hat{\theta}} (\theta - \hat{\theta})^2 \\ &= l(\mathbf{y}|\hat{\theta}) - \frac{1}{2} N \left[-\bar{l}''\big|_{\hat{\theta}}\right] (\theta - \hat{\theta})^2 \\ &= l(\mathbf{y}|\hat{\theta}) - \frac{N}{2\sigma^2} (\theta - \hat{\theta})^2 \end{aligned}$$

where $\left.\frac{dl(\mathbf{y}|\theta)}{d\theta}\right|_{\hat{\theta}} = 0$, $\bar{l}'' \equiv \frac{1}{N} \sum_{i=1}^{N} \left.\frac{d^2 l(y_i|\theta)}{d\theta^2}\right|_{\hat{\theta}}$ and $\sigma^2 := \left[-\bar{l}''\big|_{\hat{\theta}}\right]^{-1}$.[8] Then,

[7] Take into account that in the likelihood function the argument is θ. However, we keep the notation for facility in exposition.

[8] The last definition follows from standard theory in maximum likelihood estimation.

$$\begin{aligned}\pi(\theta|\mathbf{y}) &\propto \exp\{l(\mathbf{y}|\theta)\} \times \pi(\theta) \\ &\approx \exp\left\{l(\mathbf{y}|\hat{\theta}) - \frac{N}{2\sigma^2}(\theta-\hat{\theta})^2\right\} \times \pi(\theta) \\ &\propto \exp\left\{-\frac{N}{2\sigma^2}(\theta-\hat{\theta})^2\right\} \times \pi(\theta)\end{aligned}$$

Observe that we have that the posterior density is proportional to the kernel of a normal density with mean $\hat{\theta}$ and variance σ^2/N as long as $\pi(\hat{\theta}) \neq 0$. This kernel dominates as the sample size gets large due to N in the exponential term. Observe that the prior should not exclude values of θ that are logically possible, such as $\hat{\theta}$.

1.2.1 Example: Health insurance

Suppose that you are analyzing to buy a health insurance next year. To make a better decision you want to know *what is the probability that you visit your Doctor at least once next year?* To answer this question you have records of the number of times that you have visited your Doctor the last 5 years, $\mathbf{y} = \{0, 3, 2, 1, 0\}$. How to proceed?

Assuming that this is a random sample[9] from a data generating process (statistical model) that is Poisson, that is, $Y_i \sim P(\lambda)$, and your probabilistic prior beliefs about λ are well described by a Gamma distribution with shape and scale parameters α_0 and β_0, $\lambda \sim G(\alpha_0, \beta_0)$, then, you are interested in calculating the probability $P(Y_0 > 0|\mathbf{y})$. You need to calculate the posterior predictive density $\pi(Y_0|\mathbf{y})$ to answer this question in a Bayesian way.

In this example, $p(\mathbf{y}|\lambda)$ is Poisson, and $\pi(\lambda)$ is Gamma. Then, using 1.9

$$\pi(Y_0|\mathbf{y}) = \int_0^\infty \frac{\lambda^{y_0}\exp\{-\lambda\}}{y_0!} \times \pi(\lambda|\mathbf{y})d\lambda,$$

where the posterior distribution is $\pi(\lambda|\mathbf{y}) \propto \lambda^{\sum_{i=1}^N y_i+\alpha_0-1}\exp\left\{-\lambda\left(\frac{\beta_0 N+1}{\beta_0}\right)\right\}$ by equation 1.3.

Observe that the last expression is the kernel of a Gamma distribution with parameters $\alpha_n = \sum_{i=1}^N y_i + \alpha_0$ and $\beta_n = \frac{\beta_0}{\beta_0 N+1}$. Given that $\int_0^\infty \pi(\lambda|\mathbf{y})d\lambda = 1$, then the constant of proportionality in the last expression is $\Gamma(\alpha_n)\beta_n^{\alpha_n}$, where $\Gamma(\cdot)$ is the gamma function. Thus, the posterior density function $\pi(\lambda|\mathbf{y})$ is $G(\alpha_n, \beta_n)$.

Observe that

[9] Independent and identically distributed draws.

$$\begin{aligned}\mathbb{E}[\lambda|\mathbf{y}] &= \alpha_n\beta_n \\ &= \left(\sum_{i=1}^{N} y_i + \alpha_0\right)\left(\frac{\beta_0}{\beta_0 N + 1}\right) \\ &= \bar{y}\left(\frac{N\beta_0}{N\beta_0 + 1}\right) + \alpha_0\beta_0\left(\frac{1}{N\beta_0 + 1}\right) \\ &= w\bar{y} + (1-w)\mathbb{E}[\lambda],\end{aligned}$$

where $\bar{y}$ is the sample mean, which is the maximum likelihood estimator of λ in this example, $w = \left(\frac{N\beta_0}{N\beta_0+1}\right)$ and $\mathbb{E}[\lambda] = \alpha_0\beta_0$ is the prior mean. The posterior mean is a weighted average of the maximum likelihood estimator (sample information) and the prior mean. Observe that $\lim_{N\to\infty} w = 1$, that is, the sample information asymptotically dominates.

The predictive distribution is

$$\begin{aligned}\pi(Y_0|\mathbf{y}) =& \int_0^\infty \frac{\lambda^{y_0}\exp\{-\lambda\}}{y_0!} \times \frac{1}{\Gamma(\alpha_n)\beta_n^{\alpha_n}}\lambda^{\alpha_n-1}\exp\{-\lambda/\beta_n\}\,d\lambda \\ =& \frac{1}{y_0!\Gamma(\alpha_n)\beta_n^{\alpha_n}}\int_0^\infty \lambda^{y_0+\alpha_n-1}\exp\left\{-\lambda\left(\frac{1+\beta_n}{\beta_n}\right)\right\}d\lambda \\ =& \frac{\Gamma(y_0+\alpha_n)\left(\frac{\beta_n}{\beta_n+1}\right)^{y_0+\alpha_n}}{y_0!\Gamma(\alpha_n)\beta_n^{\alpha_n}} \\ =& \binom{y_0+\alpha_n-1}{y_0}\left(\frac{\beta_n}{\beta_n+1}\right)^{y_0}\left(\frac{1}{\beta_n+1}\right)^{\alpha_n}.\end{aligned}$$

The third equality follows from the kernel of a Gamma density, and the fourth from $\binom{y_0+\alpha_n-1}{y_0} = \frac{(y_0+\alpha_n-1)(y_0+\alpha_n-2)\ldots\alpha_n}{y_0!} = \frac{\Gamma(y_0+\alpha_n)}{\Gamma(\alpha_n)y_0!}$ using a property of the Gamma function.

Observe that this is a Negative Binomial density, that is $Y_0|\mathbf{y} \sim NB(\alpha_n, p_n)$ where $p_n = \frac{\beta_n}{\beta_n+1}$.

Up to this point, we have said nothing about the hyperparameters, which are required to give a concrete response to this exercise. Thus, we show two approaches to set them. First, we set $\alpha_0 = 0.001$ and $\beta_0 = 1/0.001$ which imply vague prior information about λ due to having a large degree of variability compared to the mean information.[10] In particular, $\mathbb{E}[\lambda] = 1$ and $\mathbb{V}ar[\lambda] = 1000$.

In this setting, $P(Y_0 > 0|\mathbf{y}) = 1 - P(Y_0 = 0|\mathbf{y}) \approx 0.67$. That is, the probability of visiting the Doctor at least once next year is approximately 0.67.

[10] We should be aware that there may be technical problems using this king of hyperparameters in this setting [33].

Another approach is using *Empirical Bayes*, where we set the hyperparameters maximizing the logarithm of the marginal likelihood, that is, $\left[\hat{\alpha}_0 \ \hat{\beta}_0\right]^\top = \underset{\alpha_0,\beta_0}{\text{argmax}} \ \ln p(\mathbf{y})$ where

$$\begin{aligned} p(\mathbf{y}) &= \int_0^\infty \left\{ \frac{1}{\Gamma(\alpha_0)\beta_0^{\alpha_0}} \lambda^{\alpha_0 - 1} \exp\{-\lambda/\beta_0\} \prod_{i=1}^N \frac{\lambda^{y_i} \exp\{-\lambda\}}{y_i!} \right\} d\lambda \\ &= \frac{\int_0^\infty \lambda^{\sum_{i=1}^N y_i + \alpha_0 - 1} \exp\left\{-\lambda \left(\frac{\beta_0 N + 1}{\beta_0}\right)\right\} d\lambda}{\Gamma(\alpha_0)\beta_0^{\alpha_0} \prod_{i=1}^N y_i!} \\ &= \frac{\Gamma(\sum_{i=1}^N y_i + \alpha_0) \left(\frac{\beta_0}{N\beta_0 + 1}\right)^{\sum_{i=1}^N y_i} \left(\frac{1}{N\beta_0 + 1}\right)^{\alpha_0}}{\Gamma(\alpha_0) \prod_{i=1}^N y_i} \end{aligned}$$

Using the empirical Bayes approach, we get $\hat{\alpha}_0 = 51.8$ and $\hat{\beta}_0 = 0.023$, then $P(Y_0 > 0|\mathbf{y}) = 1 - P(Y_0 = 0|\mathbf{y}) \approx 0.70$.

Observe that we can calculate the posterior odds comparing the model using an Empirical Bayes prior (model 1) versus the vague prior (model 2). We assume that $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2) = 0.5$, then

$$\begin{aligned} PO_{12} &= \frac{p(\mathbf{y}|\text{Empirical Bayes})}{p(\mathbf{y}|\text{Vague prior})} \\ &= \frac{\frac{\Gamma(\sum_{i=1}^N y_i + 51.807)\left(\frac{0.023}{N\times 0.023+1}\right)^{\sum_{i=1}^N y_i}\left(\frac{1}{N\times 0.023+1}\right)^{51.807}}{\Gamma(51.807)}}{\frac{\Gamma(\sum_{i=1}^N y_i + 0.001)\left(\frac{1/0.001}{N/0.001+1}\right)^{\sum_{i=1}^N y_i}\left(\frac{1}{N/0.001+1}\right)^{0.001}}{\Gamma(0.001)}} \\ &\approx 919. \end{aligned}$$

Then, $2 \times \log(PO_{12}) = 13.64$, there is very strong evidence against the vague prior model (see Table 1.1). In particular, $\pi(\text{Empirical Bayes}|\mathbf{y}) = \frac{919}{1+919} = 0.999$ and $\pi(\text{Vague prior}|\mathbf{y}) = 1 - 0.999 = 0.001$. These probabilities can be used to perform Bayesian model average (BMA). In particular,

$$\begin{aligned} \mathbb{E}(\lambda|\mathbf{y}) &= 1.2 \times 0.999 + 1.2 \times 0.001 = 1.2 \\ Var(\lambda|\mathbf{y}) &= 0.025 \times 0.999 + 0.24 \times 0.001 \\ &\quad + (1.2 - 1.2)^2 \times 0.999 + (1.2 - 1.2)^2 \times 0.001 = 0.025. \end{aligned}$$

The BMA predictive distribution is a mix of negative binomial distributions, that is, $y_0|\mathbf{y} \sim 0.999 \times NB(57.8, 0.02) + 0.001 \times NB(6.001, 0.17)$.

R code. Health insurance, predictive distribution using vague hyperparameters

```
set.seed(010101)
y <- c(0, 3, 2, 1, 0) # Data
N <- length(y)
ProbBo <- function(y, a0, b0){
  N <- length(y)
  #sample size
  an <- a0 + sum(y)
  # Posterior shape parameter
  bn <- b0 / ((b0 * N) + 1)
  # Posterior scale parameter
  p <- bn / (bn + 1)
  # Probability negative binomial density
  Pr <- 1 - pnbinom(0, size=an,prob=(1 - p))
  # Probability of visiting the Doctor at least once next
    year
  # Observe that in R there is a slightly different
    parametrization.
  return(Pr)
}
# Using a vague prior:
a0 <- 0.001 # Prior shape parameter
b0 <- 1 / 0.001 # Prior scale parameter
PriMeanV <- a0 * b0 # Prior mean
PriVarV <- a0 * b0^2 # Prior variance
Pp <- ProbBo(y, a0 = 0.001, b0 = 1 / 0.001)
# This setting is vague prior information.
Pp
0.67
```

R code. Health insurance, predictive distribution using empirical Bayes

```
# Using Empirical Bayes
LogMgLik <- function(theta, y){
 N <- length(y)
 #sample size
 a0 <- theta[1]
 # prior shape hyperparameter
 b0 <- theta[2]
 # prior scale hyperparameter
 an <- sum(y) + a0
 # posterior shape parameter
 if(a0 <= 0 || b0 <= 0){
  #Avoiding negative values
  lnp <- -Inf
  }else{
  lnp <- lgamma(an) + sum(y)*log(b0/(N*b0+1)) - a0*log(N*b0
    +1) - lgamma(a0)
 }
 # log marginal likelihood
 return(-lnp)
}
theta0 <- c(0.01, 1/0.1)
# Initial values
control <- list(maxit = 1000)
# Number of iterations in optimization
EmpBay <- optim(theta0, LogMgLik, method = "BFGS", control =
     control, hessian = TRUE, y = y)
# Optimization
EmpBay$convergence
0
a0EB <- EmpBay$par[1]
# Prior shape using empirical Bayes
a0EB
51.81
b0EB <- EmpBay$par[2]
# Prior scale using empirical Bayes
b0EB
0.023
PriMeanEB <- a0EB * b0EB
# Prior mean
PriVarEB <- a0EB * b0EB^2
# Prior variance
PpEB <- ProbBo(y, a0 = a0EB, b0 = b0EB)
# This setting is using emprical Bayes.
PpEB
0.70
```

R code. Health insurance, density plots

```
# Density figures:
# This code helps plotting densities
lambda <- seq(0.01, 10, 0.01)
# Values of lambda
VaguePrior <- dgamma(lambda,shape=a0,scale = b0)
EBPrior <- dgamma(lambda,shape=a0EB,scale = b0EB)
PosteriorV <- dgamma(lambda, shape = a0 + sum(y), scale = b0
    / ((b0 * N) + 1))
PosteriorEB <- dgamma(lambda, shape = a0EB+sum(y), scale =
    b0EB / ((b0EB * N) + 1))
# Likelihood function
Likelihood <- function(theta, y){
 LogL <- dpois(y, theta, log = TRUE)
 Lik <- prod(exp(LogL))
 return(Lik)
}
Liks <- sapply(lambda, function(par) {Likelihood(par, y = y)
    })
Sc <- max(PosteriorEB)/max(Liks)
#Scale for displaying in figure
LiksScale <- Liks * Sc
data <- data.frame(cbind(lambda, VaguePrior, EBPrior,
    PosteriorV, PosteriorEB, LiksScale)) #Data frame
require(ggplot2) # Cool figures
require(latex2exp) # LaTeX equations in figures
require(ggpubr) # Multiple figures in one page
fig1 <- ggplot(data = data, aes(lambda, VaguePrior)) + geom_
    line() +  xlab(TeX("$\\lambda$")) + ylab("Density") +
    ggtitle("Prior: Vague Gamma")
fig2 <- ggplot(data = data, aes(lambda, EBPrior)) + geom_
    line() + xlab(TeX("$\\lambda$")) + ylab("Density") +
    ggtitle("Prior: Empirical Bayes Gamma")
fig3 <- ggplot(data = data, aes(lambda, PosteriorV)) + geom_
    line() + xlab(TeX("$\\lambda$")) + ylab("Density") +
    ggtitle("Posterior: Vague Gamma")
fig4 <- ggplot(data = data, aes(lambda, PosteriorEB)) + geom
    _line() + xlab(TeX("$\\lambda$")) + ylab("Density") +
    ggtitle("Posterior: Empirical Bayes Gamma")
FIG <- ggarrange(fig1, fig2, fig3, fig4, ncol = 2, nrow = 2)
annotate_figure(FIG, top = text_grob("Vague versus Empirical
     Bayes: Poisson-Gamma model", color = "black", face = "
    bold", size = 14))
dataNew <- data.frame(cbind(rep(lambda, 3), c(EBPrior,
    PosteriorEB, LiksScale), rep(1:3, each = 1000))) #Data
    frame
colnames(dataNew) <- c("Lambda", "Density", "Factor")
dataNew$Factor <- factor(dataNew$Factor, levels=c("1", "3",
    "2"),
labels=c("Prior", "Likelihood", "Posterior"))
ggplot(data = dataNew, aes_string(x = "Lambda", y = "Density
    ", group = "Factor")) + geom_line(aes(color = Factor)) +
     xlab(TeX("$\\lambda$")) + ylab("Density") + ggtitle("
    Prior, likelihood and posterior: Empirical Bayes Poisson
    -Gamma model") + guides(color=guide_legend(title="
    Information")) + scale_color_manual(values = c("red", "
    yellow", "blue"))
```

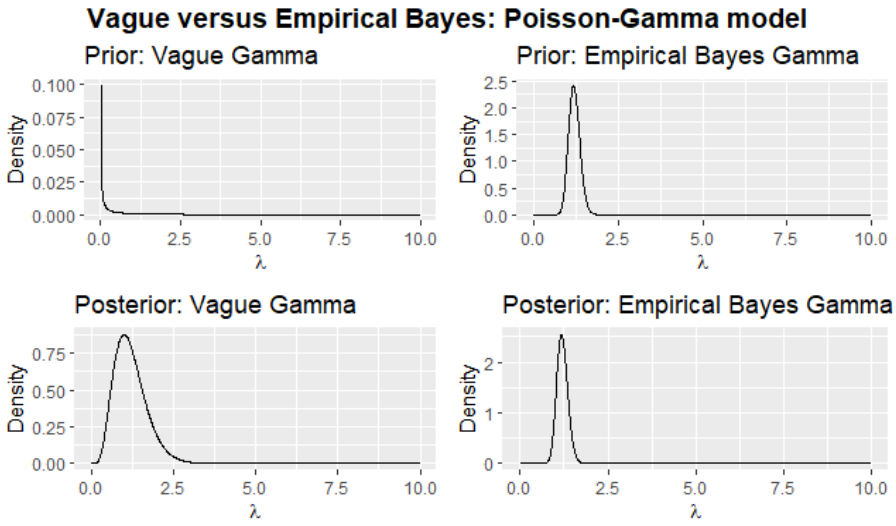

FIGURE 1.2
Vague versus Empirical Bayes: Poisson-Gamma model.

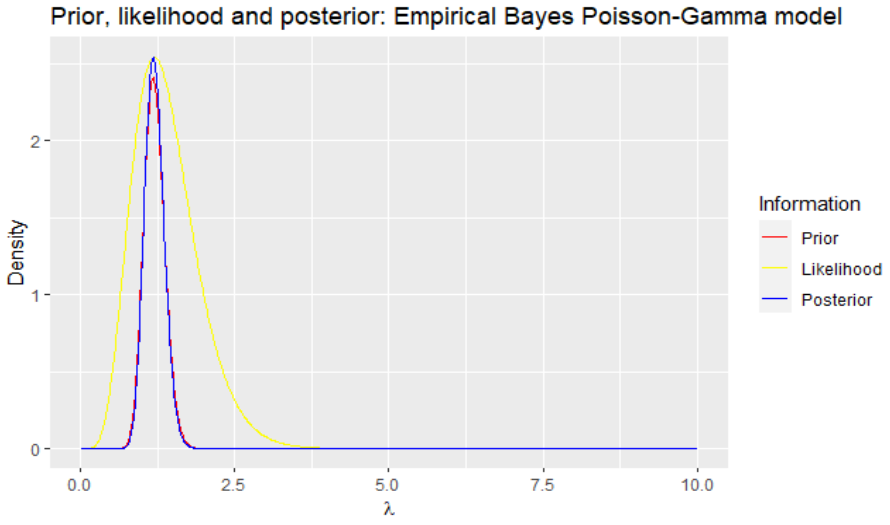

FIGURE 1.3
Prior, likelihood and posterior: Empirical Bayes Poisson-Gamma model.

Figure 1.2 displays prior and posterior densities based on vague and Empirical Bayes hyperparameters. We see that prior and posterior densities using the latter are more informative as expected.

Figure 1.3 shows the prior, scaled likelihood and posterior densities of λ based on the hyperparameters of the Empirical Bayes approach. The posterior density is a compromise between prior and sample information.

R code. Health insurance, Predictive density

```
# Predictive distributions
PredDen <- function(y, y0, a0, b0){
  N <- length(y)
  #sample size
  an <- a0 + sum(y)
  # Posterior shape parameter
  bn <- b0 / ((b0 * N) + 1)
  # Posterior scale parameter
  p <- bn / (bn + 1)
  # Probability negative binomial density
  Pr <- dnbinom(y0, size=an, prob=(1 - p))
  # Predictive density
  # Observe that in R there is a slightly different
    parametrization.
  return(Pr)
}
y0 <- 0:10
PredVague <- PredDen(y=y, y0=y0, a0=a0, b0=b0)
PredEB <- PredDen(y=y, y0=y0, a0=a0EB, b0=b0EB)
dataPred <- as.data.frame(cbind(y0, PredVague, PredEB))
colnames(dataPred) <- c("y0", "PredictiveVague", "
    PredictiveEB")
ggplot(data = dataPred) + geom_point(aes(y0, PredictiveVague
    , color = "red")) +
xlab(TeX("$y_0$")) + ylab("Density") + ggtitle("Predictive
    density: Vague and Empirical Bayes priors") + geom_point
    (aes(y0, PredictiveEB, color = "yellow")) +
guides(color = guide_legend(title="Prior")) + scale_color_
    manual(labels = c("Vague", "Empirical Bayes"), values =
    c("red", "yellow")) + scale_x_continuous(breaks=seq
    (0,10,by=1))
```

Figure 1.4 displays the predictive probability mass of not having any visits to a physician the next year, having one, two, and so on using Empirical Bayes and vague hyperparameters. The predictive probability of not having any visits are approximately equal to 30% and 33% based on the Empirical Bayes and vague hyperparameters.

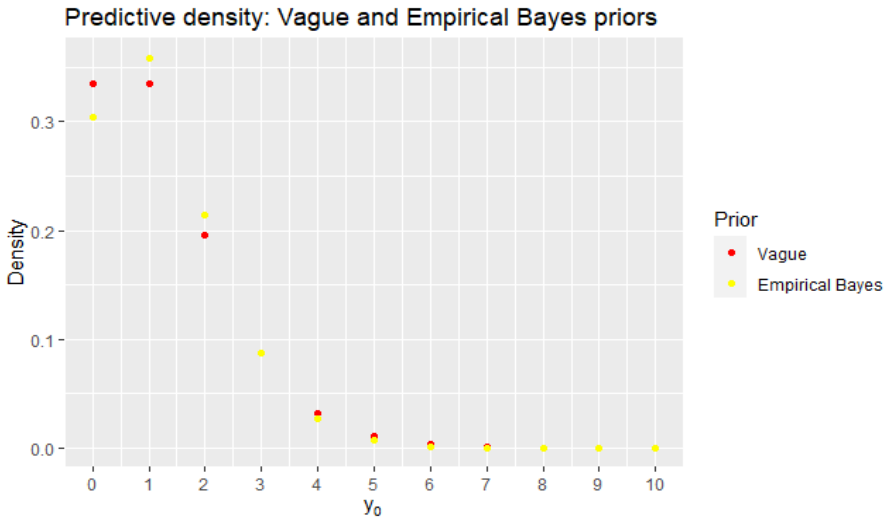

FIGURE 1.4
Predictive density: Vague and Empirical Bayes.

R code. Health insurance, Bayesian model average

```
# Posterior odds: Vague vs Empirical Bayes
PO12 <- exp(-LogMgLik(c(a0EB, b0EB), y = y))/exp(-LogMgLik(c
    (a0, b0), y = y))
PO12
919
PostProMEM <- PO12/(1 + PO12)
PostProMEM
0.998
# Posterior model probability Empirical Bayes
PostProbMV <- 1 - PostProMEM
PostProbMV
0.002
# Posterior model probability vague prior
# Bayesian model average (BMA)
PostMeanEB <- (a0EB + sum(y)) * (b0EB / (b0EB * N + 1))
# Posterior mean Empirical Bayes
PostMeanV <- (a0 + sum(y)) * (b0 / (b0 * N + 1))
# Posterior mean vague priors
BMAmean <- PostProMEM * PostMeanEB + PostProbMV * PostMeanV
BMAmean
1.2
# BMA posterior mean
PostVarEB <- (a0EB + sum(y)) * (b0EB/(b0EB * N + 1))^2
# Posterior variance Empirical Bayes
PostVarV <- (a0 + sum(y)) * (b0 / (b0 * N + 1))^2
# Posterior variance vague prior
BMAVar <- PostProMEM * PostVarEB + PostProbMV*PostVarV +
    PostProMEM * (PostMeanEB - BMAmean)^2 + PostProbMV * (
    PostMeanV - BMAmean)^2
# BMA posterior variance
BMAVar
0.025
```

R code. Health insurance, Bayesian model average

```
# BMA: Predictive
BMAPred <- PostProMEM * PredEB+PostProbMV * PredVague
dataPredBMA <- as.data.frame(cbind(y0, BMAPred))
colnames(dataPredBMA) <- c("y0", "PredictiveBMA")
ggplot(data = dataPredBMA) + geom_point(aes(y0,
    PredictiveBMA, color = "red")) +  xlab(TeX("$y_0$")) +
    ylab("Density") + ggtitle("Predictive density: BMA") +
    guides(color = guide_legend(title="BMA")) + scale_color_
    manual(labels = c("Probability"), values = c("red")) +
    scale_x_continuous(breaks=seq(0,10,by=1))

```

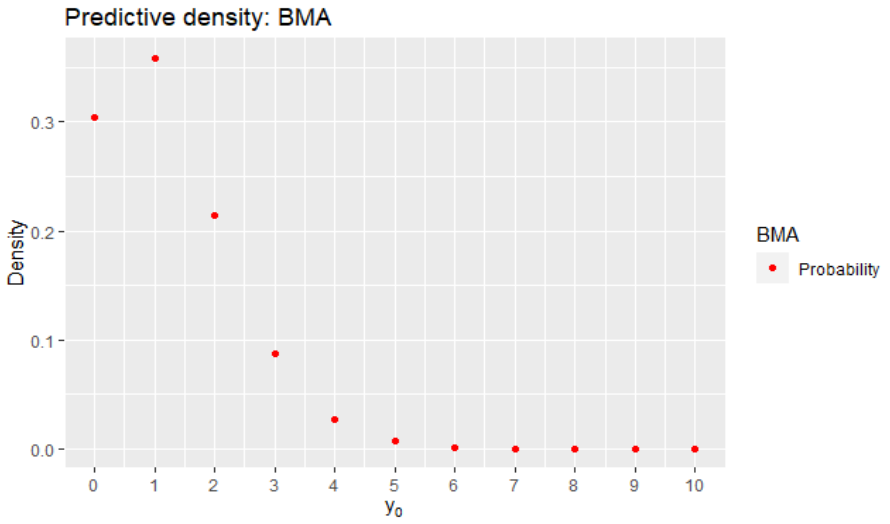

FIGURE 1.5
Bayesian model average: Predictive density.

Figure 1.5 displays the predictive density using Bayesian model average based on the vague and Empirical Bayes hyperparameters. This figure essentially resembles the predictive probability mass function based on the Empirical Bayes framework, as the posterior model probability for that setting is nearly one.

Figure 1.6 displays how the posterior distribution updates given new sample information based on an initial non-informative prior (iteration 1). We

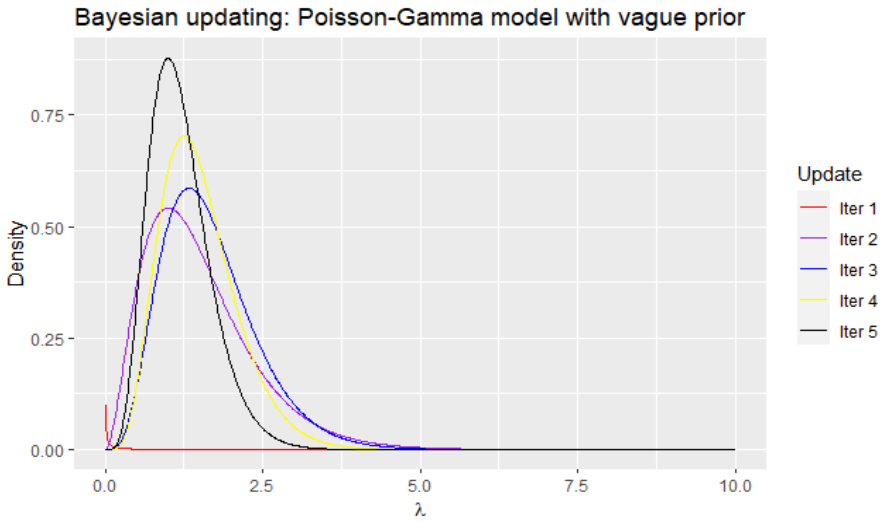

FIGURE 1.6
Bayesian updating: Posterior densities.

see that iteration 5 is based on all the sample information in our example, as a consequence, the posterior density in iteration 5 is equal to the posterior density in Figure 1.3.

R code. Health insurance, Bayes updating

```
# Bayesian updating
BayUp <- function(y, lambda, a0, b0){
  N <- length(y)
  #sample size
  an <- a0 + sum(y)
  # Posterior shape parameter
  bn <- b0 / ((b0 * N) + 1)
  # Posterior scale parameter
  p <- dgamma(lambda, shape = an, scale = bn)
  # Posterior density
  return(list(Post = p, a0New = an, b0New = bn))
}
PostUp <- NULL
for(i in 1:N){
  if(i == 1){
    PostUpi <- BayUp(y[i], lambda, a0 = 0.001, b0 = 1/0.001)
    }
  else{
    PostUpi <- BayUp(y[i], lambda, a0 = PostUpi$a0New, b0 =
    PostUpi$b0New)
  }
  PostUp <- cbind(PostUp, PostUpi$Post)
}
DataUp <- data.frame(cbind(rep(lambda, 5), c(PostUp), rep
    (1:5, each = 1000))) #Data frame
colnames(DataUp) <- c("Lambda", "Density", "Factor")
DataUp$Factor <- factor(DataUp$Factor, levels=c("1", "2", "3
    ", "4", "5"),
labels=c("Iter 1", "Iter 2", "Iter 3", "Iter 4", "Iter 5"))
ggplot(data = DataUp, aes_string(x = "Lambda", y = "Density"
    , group = "Factor")) + geom_line(aes(color = Factor)) +
    xlab(TeX("$\\lambda$")) + ylab("Density") + ggtitle("
    Bayesian updating: Poisson-Gamma model with vague prior"
    ) + guides(color=guide_legend(title="Update")) + scale_
    color_manual(values = c("red", "purple", "blue", "yellow
    ", "black"))
S <- 100000 # Posterior draws
PostMeanLambdaUps <- sapply(1:N, function(i) {mean(sample(
    lambda, S, replace = TRUE, prob = PostUp[ , i]))}) #
    Posterior mean update i
paste("Posterior means using all information and sequential
    updating are:", round(PostMeanV, 2), "and", round(
    PostMeanLambdaUps[5], 2), sep = " ")
Posterior means using all information and sequential
    updating are: 1.2 and 1.2
PostVarLambdaUps <- sapply(1:N, function(i) {var(sample(
    lambda, S, replace = TRUE, prob = PostUp[ , i]))}) #
    Posterior variance update i
paste("Posterior variances using all information and
    sequential updating are:", round(PostVarV, 2), "and",
    round(PostVarLambdaUps[5], 2), sep = " ")
Posterior variances using all information and sequential
    updating are: 0.24 and 0.24
```

1.3 Bayesian reports: Decision theory under uncertainty

The Bayesian framework allows reporting the full posterior distributions. However, some situations demand to report a specific value of the posterior distribution (point estimate), an informative interval (set), point or interval predictions and/or selecting a specific model. Decision theory offers an elegant framework to make a decision regarding what are the optimal posterior values to report [10].

The point of departure is a *loss function*, which is a non-negative real value function whose arguments are the unknown *state of nature* ($\boldsymbol{\Theta}$), and a set of *actions* to be made ($\mathcal{A}$), that is,

$$L(\boldsymbol{\theta}, a) : \boldsymbol{\Theta} \times \mathcal{A} \to \mathcal{R}^+.$$

This function is a mathematical expression of the loss of making mistakes. In particular, selecting action $a \in \mathcal{A}$ when $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is the true. In our case, the unknown state of nature can be parameters, functions of them, future or unknown realizations, models, etc.

From a Bayesian perspective, we should choose the action that minimizes the posterior expected loss ($a^*(\mathbf{y})$), that is, the *posterior risk function* ($\mathbb{E}[L(\boldsymbol{\theta}, a)|\mathbf{y}]$),

$$a^*(\mathbf{y}) = \underset{a \in \mathcal{A}}{\operatorname{argmin}}\ \mathbb{E}[L(\boldsymbol{\theta}, a)|\mathbf{y}],$$

where $\mathbb{E}[L(\boldsymbol{\theta}, a)|\mathbf{y}] = \int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta}, a)\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$.[11]

Different loss functions imply different optimal decisions. We illustrate this assuming $\theta \in \mathcal{R}$.

- The quadratic loss function, $L(\theta, a) = [\theta - a]^2$, gives as optimal decision the posterior mean, $a^*(\mathbf{y}) = \mathbb{E}[\theta|\mathbf{y}]$, that is

  $$\mathbb{E}[\theta|\mathbf{y}] = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \int_{\Theta} [\theta - a]^2 \pi(\theta|\mathbf{y})d\theta.$$

  To get this results, let's use the first condition order, differentiate the risk function with respect to a, interchange differential and integral order, and set this equal to zero, $-2\int_{\Theta}[\theta - a^*]\pi(\theta|\mathbf{y})d\theta = 0$ implies that $a^* \int_{\Theta} \pi(\theta|\mathbf{y})d\theta = a^*(\mathbf{y}) = \int_{\Theta} \theta\pi(\theta|\mathbf{y})d\theta = \mathbb{E}[\theta|\mathbf{y}]$, that is, the posterior mean is the Bayesian optimal action. This means that we should report the posterior mean as a point estimate of θ when facing the quadratic loss function.

[11] [17] propose Laplace type estimators (LTE) based on the *quasi-posterior*, $p(\boldsymbol{\theta}) = \frac{\exp\{L_n(\boldsymbol{\theta})\}\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \exp\{L_n(\boldsymbol{\theta})\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$ where $L_n(\boldsymbol{\theta})$ is not necessarily a log-likelihood function. The LTE minimizes the *quasi-posterior risk*.

- The generalized quadratic loss function, $L(\theta, a) = w(\theta)[\theta - a]^2$, where $w(\theta) > 0$ is a weighting function, gives as optimal decision rule the weighted mean. We should follow same steps as the previous result to get $a^*(\mathbf{y}) = \frac{\mathbb{E}[w(\theta)\times\theta|\mathbf{y}]}{\mathbb{E}[w(\theta)|\mathbf{y}]}$. Observe that the weighted average is driven by the weighting function $w(\theta)$.

- The absolute error loss function, $L(\theta, a) = |\theta - a|$, gives as optimal action the posterior median (Exercise 5).

- The generalized absolute error function,

$$L(\theta, a) = \begin{Bmatrix} K_0(\theta - a), \theta - a \geq 0 \\ K_1(a - \theta), \theta - a < 0 \end{Bmatrix}, \ K_0, K_1 > 0,$$

  implies the following risk function,

$$\mathbb{E}[L(\theta, a)|\mathbf{y}] = \int_{-\infty}^{a} K_1(a - \theta)\pi(\theta|\mathbf{y})d\theta + \int_{a}^{\infty} K_0(\theta - a)\pi(\theta|\mathbf{y})d\theta.$$

  Differentiating with respect to a, interchanging differentials and integrals, and equating to zero,

$$K_1 \int_{-\infty}^{a^*} \pi(\theta|\mathbf{y})d\theta - K_0 \int_{a^*}^{\infty} \pi(\theta|\mathbf{y})d\theta = 0,$$

  then, $\int_{-\infty}^{a^*} \pi(\theta|\mathbf{y})d\theta = \frac{K_0}{K_0+K_1}$, that is, any $K_0/(K_0 + K_1)$-percentile of $\pi(\theta|\mathbf{y})$ is an optimal Bayesian estimate of θ.

We can also use decision theory under uncertainty in hypothesis testing. In particular, testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, $\Theta = \Theta_0 \cup \Theta_1$ and $\emptyset = \Theta_0 \cap \Theta_1$, there are two actions of interest, a_0 and a_1, where a_j denotes no rejecting H_j, $j = \{0, 1\}$.

Given the $0 - K_j$ loss function,

$$L(\theta, a_j) = \begin{Bmatrix} 0, & \theta \in \Theta_j \\ K_j, & \theta \in \Theta_j, j \neq i \end{Bmatrix},$$

where there is no loss if the right decision is made, for instance, no rejecting H_0 when $\theta \in \Theta_0$, and the loss is K_j when an error is made, for instance, type I error, rejecting the null hypothesis (H_0) when it is true ($\theta \in \Theta_0$), implies a loss equal to K_1 due to picking a_1, no rejecting H_1.

The posterior expected loss associated with decision a_j, that is, no rejecting H_j, is $\mathbb{E}[L(\theta, a_j)|\mathbf{y}] = 0 \times P(\Theta_j|\mathbf{y}) + K_j P(\Theta_i|\mathbf{y}) = K_j P(\Theta_i|\mathbf{y})$, $j \neq i$. Therefore, the Bayes optimal decision is the one that gives the smallest posterior expected loss, that is, the null hypothesis is rejected (a_1 is not rejected), when

$K_0 P(\Theta_1|\mathbf{y}) > K_1 P(\Theta_0|\mathbf{y})$. Given our framework ($\Theta = \Theta_0 \cup \Theta_1, \emptyset = \Theta_0 \cap \Theta_1$), then $P(\Theta_0|\mathbf{y}) = 1 - P(\Theta_1|\mathbf{y})$, and as a consequence, $P(\Theta_1|\mathbf{y}) > \frac{K_1}{K_1+K_0}$, that is, the rejection region of the Bayesian test is $R = \left\{\mathbf{y} : P(\Theta_1|\mathbf{y}) > \frac{K_1}{K_1+K_0}\right\}$.

Decision theory also helps to construct interval (region) estimates. Let $\Theta_{C(\mathbf{y})} \subset \Theta$ a *credible set* for θ, and $L(\theta, \Theta_{C(\mathbf{y})}) = 1 - \mathbb{1}\left\{\theta \in \Theta_{C(\mathbf{y})}\right\}$, where

$$\mathbb{1}\left\{\theta \in \Theta_{C(\mathbf{y})}\right\} = \begin{Bmatrix} 1, & \theta \in \Theta_{C(\mathbf{y})} \\ 0, & \theta \notin \Theta_{C(\mathbf{y})} \end{Bmatrix}.$$

Then,

$$L(\theta, \Theta_{C(\mathbf{y})}) = \begin{Bmatrix} 0, & \theta \in \Theta_{C(\mathbf{y})} \\ 1, & \theta \notin \Theta_{C(\mathbf{y})} \end{Bmatrix},$$

where the 0–1 loss function is equal to zero if $\theta \in \Theta_{C(\mathbf{y})}$, and one if $\theta \notin \Theta_{C(\mathbf{y})}$. Then, the risk function is $1 - P(\theta \in \Theta_{C(\mathbf{y})})$.

Given a *measure of credibility* ($\alpha(\mathbf{y})$) that defines the level of trust that $\theta \in \Theta_{C(\mathbf{y})}$; then, we can measure the accuracy of the report by $L(\theta, \alpha(\mathbf{y})) = [\mathbb{1}\left\{\theta \in \Theta_{C(\mathbf{y})}\right\} - \alpha(\mathbf{y})]^2$. This loss function could be used to suggest a choice of the report $\alpha(\mathbf{y})$. Given that this is a quadratic loss function, the optimal action is the posterior mean, that is $\mathbb{E}[\mathbb{1}\left\{\theta \in \Theta_{C(\mathbf{y})}\right\}|\mathbf{y}] = P(\theta \in \Theta_{C(\mathbf{y})}|\mathbf{y})$. This probability can be calculated given the posterior distribution, that is, $P(\theta \in \Theta_{C(\mathbf{y})}|\mathbf{y}) = \int_{\Theta_{C(\mathbf{y})}} \pi(\theta|\mathbf{y}) d\theta$. This is a measure of the belief that $\theta \in \Theta_{C(\mathbf{y})}$ given the prior beliefs and sample information.

The set $\Theta_{C(\mathbf{y})} \in \Theta$ is a $100(1-\alpha)\%$ credible set with respect to $\pi(\theta|\mathbf{y})$ if $P(\theta \in \Theta_{C(\mathbf{y})}|\mathbf{y}) = \int_{\Theta_{C(\mathbf{y})}} \pi(\theta|\mathbf{y}) = 1 - \alpha$.

Two alternatives to report credible sets are the *symmetric credible set* and the *highest posterior density set* (HPD). The former is based on $\frac{\alpha}{2}\%$ and $(1 - \frac{\alpha}{2})\%$ percentiles of the posterior distribution, and the latter is a $100(1-\alpha)\%$ credible interval for θ with the property that it has the smallest distance compared to any other $100(1-\alpha)\%$ credible interval for θ based on the posterior distribution. That is, $C(\mathbf{y}) = \{\theta : \pi(\theta|\mathbf{y}) \geq k(\alpha)\}$, where $k(\alpha)$ is the largest number such that $\int_{\theta:\pi(\theta|\mathbf{y})\geq k(\alpha)} \pi(\theta|\mathbf{y}) d\theta = 1-\alpha$. The HPD set can be a collection of disjoint intervals when working with multimodal posterior densities. In addition, they have the limitation of not necessary being invariant under transformations.

Decision theory can also be used to perform prediction (point, sets or probabilistic). Suppose that there is a loss function $L(Y_0, a)$ involving the prediction of Y_0. Then, $\mathbb{E}_{Y_0}[L(Y_0, a)] = \int_{\mathcal{Y}_0} L(Y_0, a)\pi(Y_0|\mathbf{y}) dY_0$, where $\pi(Y_0|\mathbf{y})$ is the predictive density function. Thus, we make an optimal choice for prediction that minimizes the risk function given a specific loss function.

Although BMA allows incorporating model uncertainty in a regression framework, sometimes it is desirable to select just one model. A compelling alternative is the model with the highest posterior model probability. This model is the best alternative for prediction in the case of a 0–1 loss function [21].

1.3.1 Example: Health insurance continues

We show some optimal rules in the health insurance example. In particular, the best point estimates of λ given the quadratic, absolute and generalized absolute loss functions. For the third, we assume that underestimating λ is twice as costly as overestimating it, that is, $K_0 = 2$ and $K_1 = 1$.

Taking into account that the posterior distribution of λ is $G(\alpha_0 + \sum_{i=1}^{N} y_i, \beta_0/(\beta_0 N + 1))$, using the hyperparameters from empirical Bayes, we have that $\mathbb{E}[\lambda|\mathbf{y}] = \alpha_n\beta_n = 1.2$, the median is 1.19, and the 2/3-th quantile is 1.26. Those are the optimal point estimates for the quadratic, absolute and generalized absolute loss functions.

In addition, we test the null hypothesis H_0. $\lambda \in [0, 1)$ versus H_1. $\lambda \in [1, \infty)$ setting $K_0 = K_1 = 1$ we should reject the null hypothesis due to $P(\lambda \in [0, 1)) = 0.9 > K_1/(K_0 + K_1) = 0.5$.

We get that the 95% symmetric credible interval is (0.91, 1.53), and the highest posterior density interval is (0.9, 1.51). Finally, the optimal point prediction under a quadratic loss function is 1.2, which is the mean value of the posterior predictive distribution, and the optimal model assuming a 0-1 loss function is the model using the hyperparameters from the empirical Bayes procedure due to the posterior model probability of this model being approximately 1, whereas the posterior model probability of the model using vague hyperparameters is approximately 0.

R code. Health insurance, Bayesian reports

```
an <- sum(y) + a0EB
# Posterior shape parameter
bn <- b0EB / (N*b0EB + 1)
# Posterior scale parameter
S <- 1000000
# Number of posterior draws
Draws <- rgamma(1000000, shape = an, scale = bn)
# Posterior draws
###### Point estimation ########
OptQua <- an*bn
# Mean: Optimal choice quadratic loss function
OptQua
1.200952
OptAbs <- qgamma(0.5, shape = an, scale = bn)
# Median: Optimal choice absolute loss function
OptAbs
1.194034
# Setting K0 = 2 and K1 = 1, that is, to underestimate
    lambda is twice as costly as to overestimate it.
K0 <- 2; K1 <- 1
OptGenAbs <- quantile(Draws, K0/(K0 + K1))
# Median: Optimal choice generalized absolute loss function
OptGenAbs
66.66667%
1.262986
###### Hypothesis test ########
# H0: lambda in [0,1) vs H1: lambda in [1, Inf]
K0 <- 1; K1 <- 1
ProbH0 <- pgamma(1, shape = an, scale = bn)
ProbH0 # Posterior  probability H0
0.09569011
ProbH1 <- 1 -ProbH0
ProbH1 # Posterior  probability H1
0.9043099
# We should reject H0 given ProbH1 > K1 / (K0 + K1)
###### Credible intervals ########
LimInf <- qgamma(0.025, shape = an, scale = bn) # Lower
    bound
LimInf
0.9114851
LimSup <- qgamma(0.975, shape = an, scale = bn) # Upper
    bound
LimSup
1.529724
HDI <- HDInterval::hdi(Draws, credMass = 0.95) # Highest
    posterior density credible interval
HDI
    lower     upper
0.8971505 1.5125911
attr(,"credMass")
[1] 0.95
###### Predictive optimal choices ########
p <- bn / (bn + 1) # Probability negative binomial density
OptPred <- p/(1-p)*an # Optimal point prediction given a
    quadratic loss function in prediction
OptPred
1.200952
```

1.4 Summary

We introduce the Bayes' rule to update probabilistic statements using funny examples. Then, we study the three probabilistic objects of main relevance in Bayesian inference: the posterior distribution, the marginal likelihood and the predictive density. The first allows performing inference regarding parameters, the second is required to perform hypothesis test for model selection using the Bayes factor, and the third to perform probabilistic predictions. We also review some sampling properties of Bayesian estimators, and Bayes update. All those concepts were developed using a simple example in R software. Finally, we introduce some concepts of decision theory that can be used to report summary statistics minimizing posterior expected losses.

1.5 Exercises

1. *The court case: the blue or green cap*

   A cab was involved in a hit and run accident at night. There are two cab companies in the town: blue and green. The former has 150 cabs, and the latter 850 cabs. A witness said that a blue cab was involved in the accident; the court tested his/her reliability under the same circumstances, and got that 80% of the times the witness correctly identified the color of the cab. *What is the probability that the color of the cab involved in the accident was blue given that the witness said it was blue?*

2. *The Monty Hall problem*

   What is the probability of winning a car in the *Monty Hall problem* switching the decision if there are four doors, where there are three goats and one car? Solve this problem analytically and computationally. What if there are n doors, $n-1$ goats and one car?

3. Solve the health insurance example using a Gamma prior in the rate parametrization, that is, $\pi(\lambda) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0 - 1} \exp\{-\lambda\beta_0\}$.

4. Suppose that you are analyzing to buy a car insurance next year. To make a better decision you want to know *what is the probability that you have a car claim next year?* You have the records of your car claims in the last 15 years, $\mathbf{y} = \{0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0\}$.

   Assume that this is a random sample from a data generating process (statistical model) that is Bernoulli, $Y_i \sim Ber(p)$, and your probabilistic prior beliefs about p are well described by a beta distribution with parameters α_0 and β_0, $p \sim B(\alpha_0, \beta_0)$, then, you are

interested in calculating the probability of a claim the next year $P(Y_0 = 1|\mathbf{y})$.

Solve this using an empirical Bayes approach and a non-informative approach where $\alpha_0 = \beta_0 = 1$ (uniform distribution).

5. Show that given the loss function, $L(\theta, a) = |\theta - a|$, then the optimal decision rule minimizing the risk function, $a^*(\mathbf{y})$, is the median.

2

Conceptual differences of the Bayesian and Frequentist approaches

We give some of the conceptual differences between the Bayesian and Frequentist inferential approaches. We emphasize in the Bayesian concepts as most of the readers can be familiarized with the Frequentist statistical framework.

2.1 The concept of probability

Let's begin with the following thought experiment: Assume that you are watching the international game show "Who wants to be a millionaire?", the contestant is asked to answer a very simple question: **What is the last name of the brothers who are credited with inventing the world's first successful motor-operated airplane?**

- What is the probability that the contestant answers this question correctly?

  Unless you have:

  1. watched this particular contestant participating in this show many times,
  2. seen him asked this same question each time,
  3. and computed the relative frequency with which he gives the correct answer,

  you need to answer this question as a Bayesian!

Uncertainty about the event *answer this question* needs to be expressed as a "degree of belief" informed both by information coming from data on the skill of the particular participant, and how much he knows about inventors, and possibly prior knowledge on his performance in other game shows. Of course, your prior knowledge of the contestant may be minimal, or it may be very informed. Either way, your final answer remains a degree of belief held about an uncertain, and inherently unrepeatable state of nature.

The point of this hypothetical, light-hearted scenario is simply to highlight that a key distinction between the Frequentist and Bayesian approaches to inference is not the use (or nature) of prior information, but simply the manner

in which probability is used. To the Bayesian, probability is the mathematical construct used to quantify uncertainty about an unknown state of nature, conditional on observed data and prior knowledge about the context in which that state of nature occurs. To the Frequentist, probability is linked intrinsically to the concept of a repeated experiment, and the relative frequency with which a particular outcome occurs, conditional on that unknown state. This distinction remains key whether the Bayesian chooses to be *informative or subjective* in the specification of prior information, or chooses to be *non-informative or objective*.

Frequentists consider probability as a physical phenomenon, like mass or wavelength, whereas Bayesians stipulate that probability lives in the mind of scientists as any scientific construct [65].

It seems that the understanding of the concept of probability for the common human being is more associated with "degrees of belief" rather than relative frequency. Peter Diggle, President of The Royal Statistical Society between 2014 and 2016, was asked in an interview "A different trend which has surged upwards in statistics during Peter's career is the popularity of Bayesian statistics. Does Peter consider himself a Bayesian?", and he replied "... you can't not believe in Bayes' theorem because it's true. But that doesn't make you a Bayesian in the philosophical sense. When people are making personal decisions – even if they don't formally process Bayes' theorem in their mind – they are adapting what they think they should believe in response to new evidence as it comes in. Bayes' theorem is just the formal mathematical machinery for doing that."

However, we should say that psychological experiments suggest that human beings suffer from *anchoring*, that is, a cognitive bias that causes us to rely too heavily on the previous information (prior) such that the updating process (posterior) due to new information (likelihood) being low compared to the Bayes' rule [46].

2.2 Subjectivity is not the key

The concepts of *subjectivity* and *objectivity* indeed characterize both statistical paradigms in differing ways. Among Bayesians there are those who are immersed in *subjective* rationality [74, 23, 78, 56], but others who adopt *objective* prior distributions such as Jeffreys', reference, empirical or robust [5, 52, 44, 9] to operationalize Bayes' rule, and thereby weight quantitative (data-based) evidence. Among Frequentists, there are choices made about significance levels which, if not explicitly subjective, are typically not grounded in any objective and documented assessment of the relative losses of Type I and Type II er-

rors.[1] In addition, both Frequentist and Bayesian statisticians make decisions about the form of the data generating process, or "model", which – if not subject to rigorous diagnostic assessment – retains a subjective element that potentially influences the final inferential outcome. Although we all know that by definition a model is a schematic and simplified approximation to reality,

"Since all models are wrong the scientist cannot obtain a *correct* one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena." [13].

We also know that "All models are wrong, but some are useful" [14], that is why model diagnostics are important. This task can be performed in both approaches. Particularly, the Bayesian framework can use predictive p–values for absolute testing [30, 3] or posterior odds ratios for relative statements [43, 49]. This is because the marginal likelihood, conditional on data, is interpreted as the evidence of the prior distribution [8].

In addition, what is objectivity in a Frequentist approach? For example, why should we use a 5% or 1% significance level rather than any other value? As someone said, the apparent objectivity is really a consensus [56]. In fact "Student" (William Gosset) saw statistical significance at any level as being "nearly valueless" in itself [93]. But, this is not just a situation in the Frequentist approach. The cut–offs given to "establish" scientific evidence against a null hypothesis in terms of log_{10} scale [44] or log_e scale [49] in Table 1.1 are also *ad hoc*.

Although the true state of nature in Bayesian inference is expressed in "degrees of belief", the distinction between the two paradigms does not reside in one being more, or less, *subjective* than the other. Rather, the differences are philosophical, pedagogical, and methodological.

2.3 Estimation, hypothesis testing and prediction

All what is required to perform estimation, hypothesis testing (model selection) and prediction in the Bayesian approach is to apply the Bayes' rule. This means coherence under a probabilistic view. But, there is no free lunch, coherence reduces flexibility. On the other hand, the Frequestist approach may be not coherent from a probabilistic point of view, but it is very flexible. This approach can be seen as a tool kit that offers inferential solutions under the umbrella of understanding probability as relative frequency. For instance, a point estimator in a Frequentist approach is found such that satisfies good sampling properties like unbiasness, efficiency, or a large sample property as consistency.

[1] Type I error is rejecting the null hypothesis when this is true, and the Type II error is not rejecting the null hypothesis when this is false.

A remarkable difference is that optimal Bayesian decisions are calculated minimizing the expected value of the loss function with respect to the posterior distribution, that is, it is conditional on observed data. On the other hand, Frequentist "optimal" actions are base on the expected values over the distribution of the estimator (a function of data) conditional on the unknown parameters, that is, it considers sampling variability.

The Bayesian approach allows to obtain the posterior distribution of any unknown object such as parameters, latent variables, future or unobserved variables or models. A nice advantage is that prediction can take into account estimation error, and predictive distributions (probabilistic forecasts) can be easily recovered.

Hypothesis testing (model selection) is based on *inductive logic* reasoning (*inverse probability*); on the basis of what we see, we evaluate what hypothesis is most tenable, and is performed using posterior odds, which in turn are based on Bayes factors that evaluate evidence in favor of a null hypothesis taking explicitly the alternative [49], following the rules of probability [56], comparing how well the hypothesis predicts data [39], minimizing the weighted sum of type I and type II error probabilities [24, 66], and taking the implicit balance of losses [44, 11] into account. Posterior odds allows to use the same framework to analyze nested and non-nested models and perform model average. However, Bayes factors cannot be based on improper or vague priors [50], the practical interplay between model selection and posterior distributions is not as easy as it maybe in the Frequentist approach, and the computational burden can be more demanding due to solving potentially difficult integrals.

On the other hand, the Frequentist approach establishes most of its estimators as the solution of a system of equations. Observe that optimization problems reduce to solve systems. We can potentially get the distribution of these estimators, but most of the time it is needed asymptotic arguments or resampling techniques. Hypothesis testing requires pivotal quantities and/or also resampling, and prediction most of the time is based on a *plug-in approach*, which means not taking estimation error into account.[2] In addition, ancillary statistics can be used to build prediction intervals.[3] Comparing models depends on their structure, for instance, there are different Frequentist statistical approaches to compare nested and non-nested models. A nice feature in some situations is that there is a practical interplay between hypothesis testing and confidence intervals, for instance in the normal population mean hypothesis framework you cannot reject at α significance level (Type I error) any null hypothesis H_0. $\mu = \mu^0$ if μ^0 is in the $1-\alpha$ confidence interval $P(\mu \in [\hat{\mu} - |t_{N-1}^{\alpha/2}| \times \hat{\sigma}_{\hat{\mu}}, \hat{\mu} + |t_{N-1}|^{\alpha/2} \times \hat{\sigma}_{\hat{\mu}}]) = 1-\alpha$, where $\hat{\mu}$ and $\hat{\sigma}_{\hat{\mu}}$ are the maximum likelihood estimators of the mean and standard error, and $t_{N-1}^{\alpha/2}$ is

[2] A pivot quantity is a function of unobserved parameters and observations whose probability distribution does not depend on the unknown parameters.

[3] An ancillary statistic is a pivotal quantity that is also a statistic.

the quantile value of the Student's t distribution at $\alpha/2$ probability and $N-1$ degrees of freedom, N is the sample size.

A remarkable difference between the Bayesian and the Frequentist inferential frameworks is the interpretation of credible/confidence intervals. Observe that once we have estimates, such that for example the previous interval is $[0.2, 0.4]$ given a 95% confidence level, we cannot say that $P(\mu \in [0.2, 0.4]) = 0.95$ in the Frequentist framework. In fact, this probability is 0 or 1 under this approach, as μ can be there or not, the problem is that we will never know in applied settings. This due to that $P(\mu \in [\hat{\mu} - |t_{N-1}^{0.025}| \hat{\times} \sigma_{\hat{\mu}}, \hat{\mu} + |t_{N-1}^{0.025}| \times \hat{\sigma}_{\hat{\mu}}]) = 0.95$ being in the sense of repeated sampling. On the other hand, once we have the posterior distribution, we can say that $P(\mu \in [0.2, 0.4]) = 0.95$ under the Bayesian framework.

Following common practice, most of researchers and practitioners do hypothesis testing based on the p-value in the Frequentist framework. But, **what is a *p*–value?** Most of the users do not know the answer due to many time statistical inference is not performed by statisticians [9].[4] A p–value is the probability of obtaining a statistical summary of the data equal to or *more extreme* than what was actually observed, assuming that the null hypothesis is true.

Therefore, p–value calculations involve not just the observed data, but also more *extreme* hypothetical observations. So,

"What the use of p implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred."[44]

It seems that common Frequentist inferential practice intertwined two different logic reasoning arguments: the p–value [27] and *significance level* [64]. The former is an informal short–run criterion, whose philosophical foundation is *reduction to absurdity*, which measures the discrepancy between the data and the null hypothesis. So, the p–value is not a direct measure of the probability that the null hypothesis is false. The latter, whose philosophical foundations is *deduction*, is based on a long–run performance such that controls the overall number of incorrect inferences in the repeated sampling without care of individual cases. The p–value fallacy consists in interpreting the p–value as the strength of evidence against the null hypothesis, and using it simultaneously with the frequency of type I error under the null hypothesis [39].

The American Statistical Association has several concerns regarding the use of the p–value as a cornerstone to perform hypothesis testing in science. This concern motivates the ASA's statement on p–values [90], which can be summarized in the following principles:

- "P–values can indicate how incompatible the data are with a specified statistical model."

[4] https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/

- "P–values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone."
- "Scientific conclusions and business or policy decisions should not be based only on whether a p–value passes a specific threshold."
- "Proper inference requires full reporting and transparency."
- "A p–value, or statistical significance, does not measure the size of an effect or the importance of a result."
- "By itself, a p–value does not provide a good measure of evidence regarding a model or hypothesis."

To sum up, Fisher proposed the p-value as a witness rather than a judge. So, a p-value lower than the significance level means more inspection of the null hypothesis, but it is not a final conclusion about it.

Another difference between the Frequentists and the Bayesians is the way how scientific hypothesis are tested. The former use the p-value, whereas the latter use the Bayes factor. Observe that the p–value is associated with the probability of the data given the hypothesis, whereas the Bayes factor is associated with the probability of the hypothesis given the data. However, there is an approximate link between the t statistic and the Bayes factor for regression coefficients [70]. In particular, $|t| > (log(N)+6)^{1/2}$, corresponds to strong evidence in favor of rejecting the not relevance of a control in a regression. Observe that in this setting the threshold of the t statistic, and as a consequence the significant level, depends on the sample size. Observe that this setting agrees with the idea in experimental designs of selecting the sample size such that we control Type I and Type II errors. In observational studies we cannot control the sample size, but we can select the significance level.

See also [80] and [7] for nice exercises to reveal potential flaws of the p–value (p) due to $p \sim U[0,1]$ under the null hypothesis,[5] and calibrations of the p-value to interpret them as the odds ratio and the error probability. In particular, $B(p) = -e \times p \times \log(p)$ when $p < e^{-1}$, and interpret this as the Bayes factor of H_0 to H_1, where H_1 denotes the unspecified alternative to H_0, and $\alpha(p) = (1 + [-e \times p \times \log(p)]^{-1})^{-1}$ as the error probability α in rejecting H_0. Take into account that $B(p)$ and $\alpha(p)$ are lower bounds.

Logic of argumentation in the Frequentist approach is based on *deductive logic*, this means that it starts from a statement about the true state of nature (null hypothesis), and predicts what should be seen if this statement were true. On the other hand, the Bayesian approach is based on *inductive logic*, this means that it defines what hypothesis is more consistent with what is seen. The former inferential approach establishes that the true of the premises implies the true of the conclusion, that is why we reject or not reject hypothesis. The latter establishes that the premises supply some evidence, but not

[5] https://joyeuserrance.wordpress.com/2011/04/22/proof-that-p-values-under-the-null-are-uniformly-distributed/ for a simple proof.

full assurance, of the true of the conclusion, that is why we get probabilistic statements.

Here, there is a difference between *effects of causes* (forward causal inference) and *causes of effects* (reverse causal inference) [34, 22]. To illustrate this point, imagine that a firm increases the price of a specific good, then economic theory would say that its demand decreases. The premise (null hypothesis) is a price increase, and the consequence is a demand reduction. Another view would be to observe a demand reduction, and try to identify which cause is more tenable. For instance, demand reduction can be caused by any positive supply shocks or any negative demand shocks. The Frequentist logic sees the first view, and the Bayesian reasoning gives the probability associated with possible causes.

2.4 The likelihood principle

The **likelihood principle** states that in making inference or decisions about the state of the nature all the relevant *experimental* information is given by the *likelihood function*. The Bayesian framework follows this statement, that is, it is conditional on observed data.

We follow [8], who in turns followed [57], to illustrate the likelihood principle. We are given a coin such that we are interested in the probability, θ, of having it come up heads when flipped. It is desired to test H_0. $\theta = 1/2$ versus H_1. $\theta > 1/2$. An experiment is conducted by flipping the coin (independently) in a series of trials, the results of which is the observation of 9 heads and 3 tails.

This is not yet enough information to specify $p(y|\theta)$, since the series of trials was not explained. Two possibilities:

1. The experiment consisted of a predetermine 12 flips, so that $Y = [Heads]$ would be $B(12, \theta)$, then $p_1(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = 220 \times \theta^9(1-\theta)^3$.
2. The experiment consisted of flipping the coin until 3 tails were observed ($r = 3$). Then, Y, the number of failures (heads) until getting 3 tails, is $NB(3, 1-\theta)$. Then, $p_2(y|\theta) = \binom{y+r-1}{r-1}(1-(1-\theta)^y(1-\theta)^r = 55 \times \theta^9(1-\theta)^3$.

Using a Frequentist approach, the significance level of $y = 9$ using the Binomial model against $\theta = 1/2$ would be:

$$\alpha_1 = P_{1/2}(Y \geq 9) = p_1(9|1/2) + p_1(10|1/2) + p_1(11|1/2) + p_1(12|1/2) = 0.073.$$

R code. The likelihood principle: Binomial model

```
success <- 9
# Number of observed success in n trials
n <- 12
# Number of trials
siglevel <- sum(sapply(9:n,function(y)dbinom(y,n,0.5)))
siglevel
0.073
```

For the Negative Binomial model, the significance level would be:

$$\alpha_2 = P_{1/2}(Y \geq 9) = p_2(9|1/2) + p_2(10|1/2) + \ldots = 0.0327.$$

R code. The likelihood principle: Negative Binomial model

```
success <- 3
# Number of target success (tails)
failures <- 9
# Number of failures
siglevel <- 1 - pnbinom((failures - 1),success,0.5)
siglevel
0.0327
```

We arrive to different conclusions using a significance level equal to 5%, whereas we obtain the same outcomes using a Bayesian approach because the kernels of both distributions are the same ($\theta^9 \times (1-\theta)^3$).

2.5 Why is not the Bayesian approach that popular?

At this stage, we may wonder why the Bayesian statistical framework is not the dominant inferential approach despite that it has its historical origin in 1763 [6], whereas the Frequentist statistical framework was largely developed in the early 20th century. The scientific battle over the Bayesian inferential approach lasted for 150 years, and this maybe explained by some of the following facts.

There is an issue regarding *apparent subjectivity* as the Bayesian inferential approach runs counter the strong conviction that science demands objectivity, and Bayesian probability is a measure of degrees of belief, where the initial prior maybe just a guess; this was not accepted as objective and rigorous science. Initial critics said that Bayes was quantifying ignorance as he set equal probabilities to any potential result. As a consequence, prior distributions were damned [62].

Bayes himself seemed not to have believed in his idea. Although, it seems that Bayes achieved his breakthrough during the late 1740s, he did not send it off to the Royal Society for publication. It was his friend, Richard Price, another Presbyterian minister, who rediscovered Bayes' idea, polished it and published.

However, it was Laplace who independently generalized Bayes' theorem in 1781. He used it initially in gambling problems, and soon after in astronomy, mixing different sources of information in order to leverage research in specific situations where data was scarce. Then, he wanted to use his discovery to find the probability of causes, and thought that this required large data sets, and turned into demography. In this field, he had to perform large calculations that demanded to develop smart approximations, creating the Laplace's approximation and the central limit theorem [52]; although, at the cost of apparently leaving his research on Bayesian inference.

Once *Laplace was gone in 1827*, the Bayes' rule disappeared from the scientific spectrum for almost a century. In part, personal attacks against Laplace made the rule be forgotten, and also, the old fashion thought that statistics does not have to say anything about causation, and that the prior is very subjective to be compatible with science. Although, practitioners used it to solve problems in astronomy, communication, medicine, military and social issues with remarkable results.

Thus, the concept of degrees of belief to operationalize probability was gone in name of scientific objectivity, and probability as the frequency an event occurs in many repeatable trials became the rule. Laplace critics argued that those concepts were diametric opposites, although, Laplace considered them as basically equivalent when large sample sizes are involved [62].

The era of the Frequentists or sampling theorists began, lead by Karl Pearson, and his nemesis, Ronald Fisher, both brilliant, against the inverse probability approach, persuasive and dominant characters that made impossible to argue against their ideas. Karl Pearson legacy was taken by his son Egon, and Egon's friend Neyman, both inherited the anti-Bayesian and anti-Fisher legacy.

Despite the anti-Bayesian campaign among statisticians, there were some independent characters developing Bayesian ideas, Borel, Ramsey and de Fineti, all of them isolated in different countries, France, England and Italy. However, the anti-Bayesian trio of Fisher, Neyman and Egon Person got all the attention during the 1920s and 1930s. Only, a geophysicist, Harold Jeffreys, kept alive Bayesian inference in the 1930s and 1940s. Jeffreys was a

very quiet, shy, uncommunicative gentleman working at Cambridge in the astronomy department. He was Fisher's friend thanks to his character, although they were diametric opposites regarding the Bayesian inferential approach, facing very high intellectual battles. Unfortunately for the Bayesian approach, *Jeffreys lost*, he was very technical using confusing high level mathematics, worried about inference from scientific evidence, not guiding future actions based on decision theory, which was very important in that era for mathematical statistics due to the Second World War. On the other hand, Fisher was a very dominant character, persuasive in public and a master of practice, his techniques were written in a popular style with minimum mathematics.

However, Bayes' rule achieved remarkable results in applied settings like the AT&T company or the social security system in USA. Bayesian inference also had a relevant role during the second World War and the Cold War. Alan Turing used inverse probability at Bletchley Park to crack German messages called Enigma code used by U-boats, Andrei Kolmogorov used it to improved firing tables of Russia's artillery, Bernard Koopman applied it for searching targets in the open sea and the RAND Corporation used it in the Cold War. Unfortunately, *these Bayesian developments were top secrets for almost 40 years* that keep classified the contribution of inverse probability in modern human history.

During 1950s and 1960s three mathematicians lead the rebirth of the Bayesian approach, Good, Savage and Lindley. However, it seems that they were unwilling to apply their theories to real problems, and despite that the Bayesian approach proved its worth, for instance, in business decisions, navy search, lung cancer, etc, it was applied to simple models due to its *mathematical complexity and requirement of large computations*. But, there were some breakthrough that change this. First, hierarchical models introduced by Lindley and Smith, where a complex model is decomposed into many easy to solve models, and second, Markov chain Monte Carlo methods developed by Hastings in the 1970s [41] and the Geman brothers in the 1980s [35]. These methods were introduced into the Bayesian inferential framework in the 1990s by Gelfand and Smith [28], and Tierney [88], when desktop computers got enough computational power to solve complex models. Since then, the Bayesian inferential framework has gained increasing popularity among practitioners and scientists.

2.6 A simple working example

We will illustrate some conceptual differences between the Bayesian and Frequentist statistical approaches performing inference given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]$, where $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, N$.

In particular, we set $\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma) \propto \frac{1}{\sigma}$. This is a standard *non-*

informative improper prior (Jeffreys prior, see Chapter 4), that is, this prior is perfectly compatible with sample information. In addition, we are assuming independent priors for μ and σ. Then,

$$\begin{aligned}
\pi(\mu,\sigma|\mathbf{y}) &\propto \frac{1}{\sigma}\times(\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right\}\\
&= \frac{1}{\sigma}\times(\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}((y_i-\bar{y})-(\mu-\bar{y}))^2\right\}\\
&= \frac{1}{\sigma}\exp\left\{-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right\}\times(\sigma)^{-N}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2\right\}\\
&= \frac{1}{\sigma}\exp\left\{-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right\}\times(\sigma)^{-(\alpha_n+1)}\exp\left\{-\frac{\alpha_n\hat{\sigma}^2}{2\sigma^2}\right\},
\end{aligned}$$

where $\bar{y}=\frac{\sum_{i=1}^{N}}{N}$, $\alpha_n=N-1$ and $\hat{\sigma}^2=\frac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{N-1}$.

The first term in the last expression is the kernel of a normal density, $\mu|\sigma,\mathbf{y}\sim N(\bar{y},\sigma^2/N)$. The second term is the kernel of an inverted gamma density [92], $\sigma|\mathbf{y}\sim IG(\alpha_n,\hat{\sigma}^2)$. Therefore, $\pi(\mu|\sigma,\mathbf{y})=(2\pi\sigma^2/N)^{-1/2}\exp\left\{\frac{-N}{2\sigma^2}(\mu-\bar{y})^2\right\}$ and $\pi(\sigma|\mathbf{y})=\frac{2}{\Gamma(\alpha_n/2)}\left(\frac{\alpha_n\hat{\sigma}^2}{2}\right)^{\alpha_n/2}\frac{1}{\sigma^{\alpha_n+1}}\times\exp\left\{-\frac{\alpha_n\hat{\sigma}^2}{2\sigma^2}\right\}$.

Observe that $\mathbb{E}[\mu|\sigma,\mathbf{y}]=\bar{y}$, this is also the maximum likelihood (Frequentist) point estimate of μ in this setting. In addition, the Frequentist $(1-\alpha)\%$ confidence interval and the Bayesian $(1-\alpha)\%$ credible interval have exactly the same form, $\bar{y}\pm|z_{\alpha/2}|\frac{\sigma}{\sqrt{N}}$, where $z_{\alpha/2}$ is the $\alpha/2$ percentile of a standard normal distribution. However, the interpretations are totally different. The confidence interval has a probabilistic interpretation under sampling variability of $\bar{Y}$, that is, in repeated sampling $(1-\alpha)\%$ of the intervals $\bar{Y}\pm|z_{\alpha/2}|\frac{\sigma}{\sqrt{N}}$ would include μ, but given an observed realization of $\bar{Y}$, say $\bar{y}$, the probability of $\bar{y}\pm|z_{\alpha/2}|\frac{\sigma}{\sqrt{N}}$ including μ is 1 or 0, that is why we say a $(1-\alpha)\%$ confidence interval. On the other hand, $\bar{y}\pm|z_{\alpha/2}|\frac{\sigma}{\sqrt{N}}$ has a simple probabilistic interpretation in the Bayesian framework, there is a $(1-\alpha)\%$ probability that μ lies in this interval.

If we want to get the marginal posterior density of μ,

$$
\begin{aligned}
\pi(\mu|\mathbf{y}) &= \int_0^\infty \pi(\mu,\sigma|\mathbf{y})d\sigma \\
&\propto \int_0^\infty \frac{1}{\sigma}\times(\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^N(y_i-\mu)^2\right\}d\sigma \\
&= \int_0^\infty \left(\frac{1}{\sigma}\right)^{N+1}\exp\left\{-\frac{N}{2\sigma^2}\frac{\sum_{i=1}^N(y_i-\mu)^2}{N}\right\}d\sigma \\
&= \left[\frac{2}{\Gamma(N/2)}\left(\frac{N\sum_{i=1}^N(y_i-\mu)^2}{2N}\right)^{N/2}\right]^{-1} \\
&\propto \left[\sum_{i=1}^N(y_i-\mu)^2\right]^{-N/2} \\
&= \left[\sum_{i=1}^N((y_i-\bar{y})-(\mu-\bar{y}))^2\right]^{-N/2} \\
&= [\alpha_n\hat{\sigma}^2+N(\mu-\bar{y})^2]^{-N/2} \\
&\propto \left[1+\frac{1}{\alpha_n}\left(\frac{\mu-\bar{y}}{\hat{\sigma}/\sqrt{N}}\right)^2\right]^{-(\alpha_n+1)/2}.
\end{aligned}
$$

The fourth line is due to having the kernel of a inverted gamma density with N degrees of freedom in the integral [92].

The last expression is the kernel of a Student's t density function with $\alpha_n = N-1$ degrees of freedom, expected value equal to $\bar{y}$, and variance $\frac{\hat{\sigma}^2}{N}\left(\frac{\alpha_n}{\alpha_n-2}\right)$. Then, $\mu|\mathbf{y} \sim t\left(\bar{y}, \frac{\hat{\sigma}^2}{N}\left(\frac{\alpha_n}{\alpha_n-2}\right), \alpha_n\right)$.

Observe that a $(1-\alpha)\%$ confidence interval and $(1-\alpha)\%$ credible interval have exactly the same expression, $\bar{y} \pm |t^{\alpha_n}_{\alpha/2}|\frac{\hat{\sigma}}{\sqrt{N}}$, where $t^{\alpha_n}_{\alpha/2}$ is the $\alpha/2$ percentile of a Student's t distribution. But again, the interpretations are totally different.

The mathematical similarity between the Frequentist and Bayesian expressions in this example is due to using an improper prior.

2.6.1 Example: Math test

You have a random sample of math scores of size $N = 50$ from a normal distribution, $Y_i \sim N(\mu,\sigma)$. The sample mean and variance are equal to 102 and 10, respectively. Assuming an improper prior equal to $1/\sigma$,

- Get 95% confidence and credible intervals for μ.
- What is the posterior probability that $\mu > 103$?

Using $\mu|\mathbf{y} \sim t\left(\bar{y}, \frac{\hat{\sigma}^2}{N}\left(\frac{\alpha_n}{\alpha_n - 2}\right), \alpha_n\right)$, which implies that $\bar{y} \pm |t^{\alpha_n}_{\alpha/2}| \frac{\hat{\sigma}}{\sqrt{N}}$, where $\bar{y} = 102$, $\hat{\sigma}^2 = 10$ and $\alpha_n = 49$, the 95% confidence and credible intervals for μ are the same (101.1, 102.9), and $P(\mu > 103) = 1.49\%$ given the sample information.

R code. Example: Math test

```
N <- 50 # Sample size
y_bar <- 102 # Sample mean
s2 <- 10 # Sample variance
alpha <- N - 1
serror <- (s2/N)^0.5
LimInf <- y_bar - abs(qt(0.025, alpha)) * serror
LimInf
101.101
# Lower bound
LimSup <- y_bar + abs(qt(0.025, alpha)) * serror
LimSup
102.898
# Upper bound
y.cut <- 103
P <- 1-metRology::pt.scaled(y.cut, df = alpha, mean = y_bar,
    sd = serror)
P
0.0149
# Probability of mu greater than y.cut
```

2.7 Summary

The differences between the Bayesian and Frequentist inferential approaches are philosophical, including as pertains to the role of probability; pedagogical, in particular as relates to the use of inference to inform decision making; and methodological, as having differences in their mathematical and computational frameworks. Although at methodological level, the debate has become considerably muted, except for some aspects of inference, with the recognition that each approach has a great deal to contribute to statistical practice [38, 4, 48]. As Bradley Efron said "Computer-age statistical inference at its most successful **combines** elements of the two philosophies" [26].

2.8 Exercises

1. **Jeffreys-Lindley's paradox**

   The **Jeffreys-Lindley's paradox** [44, 58] is an apparent disagreement between the Bayesian and Frequentist frameworks to a hypothesis testing situation.

   In particular, assume that in a city 49,581 boys and 48,870 girls have been born in 20 years. Assume that the male births is distributed Binomial with probability θ. We want to test the null hypothesis H_0. $\theta = 0.5$ versus H_1. $\theta \neq 0.5$.

   - Show that the posterior model probability for the model under the null is approximately 0.95. Assume $\pi(H_0) = \pi(H_1) = 0.5$, and $\pi(\theta)$ equal to $U(0,1)$ under H_1.
   - Show that the p-value for this hypothesis test is equal to 0.0235 using the normal approximation, $Y \sim N(N\times\theta, N\times\theta\times(1-\theta))$.

2. We want to test H_0. $\mu = \mu_0$ vs H_1. $\mu \neq \mu_0$ given $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$.

   Assume $\pi(H_0) = \pi(H_1) = 0.5$, and $\pi(\mu, \sigma) \propto 1/\sigma$ under the alternative hypothesis.

   Show that

   $p(\mathbf{y}|\mathcal{M}_1) = \frac{\pi^{-N/2}}{2}\Gamma(N/2)2^{N/2}\left(\frac{1}{\alpha_n\hat{\sigma}^2}\right)^{N/2}\left(\frac{N}{\alpha_n\hat{\sigma}^2}\right)^{-1/2}\frac{\Gamma(1/2)\Gamma(\alpha_n/2)}{\Gamma((\alpha_n+1)/2)}$

   and $p(\mathbf{y}|\mathcal{M}_0) = (2\pi)^{-N/2}\left[\frac{2}{\Gamma(N/2)}\left(\frac{N}{2}\frac{\sum_{i=1}^N(y_i-\mu_0)^2}{N}\right)^{N/2}\right]^{-1}$. Then,

   $$\begin{aligned} PO_{01} &= \frac{p(\mathbf{y}|\mathcal{M}_0)}{p(\mathbf{y}|\mathcal{M}_1)} \\ &= \frac{\Gamma((\alpha_n+1)/2)}{\Gamma(1/2)\Gamma(\alpha_n/2)}(\alpha_n\hat{\sigma}^2/N)^{-1/2}\left[1+\frac{(\mu_0-\bar{y})^2}{\alpha_n\hat{\sigma}^2/N}\right]^{-\left(\frac{\alpha_n+1}{2}\right)}, \end{aligned}$$

   where $\alpha_n = N-1$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^N(y_i-\bar{y})^2}{N-1}$.

   Find the relationship between the posterior odds and the classical test statistic for the null hypothesis.

3. **Math test continues**

   Using the setting of the **Example: Math test** in subsection 2.6.1, test H_0. $\mu = \mu_0$ vs H_1. $\mu \neq \mu_0$ where $\mu_0 = \{100, 100.5, 101, 101.5, 102\}$.

   - What is the p-value for these hypothesis tests?

- Find the posterior model probability of the null model for each μ_0.

3

Objective and subjective Bayesian approaches

4

Cornerstone models: Conjugate families

We will introduce conjugate families in basic statistical models with examples, solving them analytically and computationally using R. We will have some mathematical, and computational exercises in R.

4.1 Motivation of conjugate families

Observing the three fundamental pieces of Bayesian analysis: the posterior distribution (parameter inference), the marginal likelihood (hypothesis testing), and the predictive distribution (prediction), equations 4.1, 4.2 and 4.3, respectively,

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{4.1}$$

$$p(\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{4.2}$$

and

$$p(\mathbf{Y}_0|\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{Y}_0|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \tag{4.3}$$

we can understand that some of the initial limitations of the application of the Bayesian analysis were associated with the absence of algorithms to draw from non-standard posterior distributions (equation 4.1), and the lack of analytical solutions of the marginal likelihood (equation 4.2) and the predictive distribution (equation 4.3). Both issues requiring computational power.

Although there were algorithms to sample from non-standard posterior distributions since the second half of the last century [63, 41, 35], their particular application in the Bayesian framework emerged later [28, 88], maybe until the increasing computational power of desktop computers. However, it is also common practice nowadays to use models that have standard conditional posterior distributions to mitigate computational requirements. In addition, nice mathematical tricks plus computational algorithms [29, 19, 20] and approximations [89, 45] are used to obtain the marginal likelihood (prior predictive).

Despite these advances, there are two potentially conflicting desirable model specification features that we can see from equations 4.1, 4.2 and 4.3: (1) analytical solutions and (2) the posterior distribution in the same family as the prior distribution for a given likelihood. The latter is called *conjugate priors*, a family of priors that is closed under sampling [79].

These features are desirable as the former implies facility to perform hypothesis testing and predictive analysis, and the latter means invariance of the prior-to-posterior updating. Both features imply less computational burden.

We can easily achieve each of these features independently, for instance using improper priors for analytical tractability, and defining in a broad sense the family of prior distributions for prior conjugacy. However, these features are in conflict.

Fortunately, we can achieve these two nice characteristics if we assume that the data generating process is given by a distribution function in the *exponential family*. That is, given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$, a probability density function $p(\mathbf{y}|\boldsymbol{\theta})$ belongs to the exponential family if it has the form

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\theta}) &= \prod_{i=1}^{N} h(y_i) C(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\theta})^\top \mathbf{T}(y_i)\right\} \\ &= h(\mathbf{y}) C(\boldsymbol{\theta})^N \exp\left\{\eta(\boldsymbol{\theta})^\top \mathbf{T}(\mathbf{y})\right\} \\ &= h(\mathbf{y}) \exp\left\{\eta(\boldsymbol{\theta})^\top \mathbf{T}(\mathbf{y}) - A(\boldsymbol{\theta})\right\}, \end{aligned} \tag{4.4}$$

where $h(\mathbf{y}) = \prod_{i=1}^{N} h(y_i)$ is a non-negative function, $\eta(\boldsymbol{\theta})$ is a known function of the parameters, $A(\boldsymbol{\theta}) = \log\left\{\int_{\mathbf{Y}} h(\mathbf{y}) \exp\left\{\eta(\boldsymbol{\theta})^\top \mathbf{T}(\mathbf{y})\right\} d\mathbf{y}\right\} = -N \log(C(\boldsymbol{\theta}))$ is a normalization factor, and $\mathbf{T}(\mathbf{y}) = \sum_{i=1}^{N} \mathbf{T}(y_i)$ is the vector of sufficient statistics of the distribution (by the factorization theorem).

If the support of $\mathbf{y}$ is independent of $\boldsymbol{\theta}$, then the family is said to be *regular*, otherwise it is irregular. In addition, if we set $\eta = \eta(\boldsymbol{\theta})$, then the exponential family is said to be in the *canonical form*

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\theta}) &= h(\mathbf{y}) D(\boldsymbol{\eta})^N \exp\left\{\eta^\top \mathbf{T}(\mathbf{y})\right\} \\ &= h(\mathbf{y}) \exp\left\{\eta^\top \mathbf{T}(\mathbf{y}) - B(\boldsymbol{\eta})\right\}. \end{aligned}$$

A nice feature of this representation is that $\mathbb{E}[\mathbf{T}(\mathbf{y})|\boldsymbol{\eta}] = \nabla B(\boldsymbol{\eta})$ and $Var[\mathbf{T}(\mathbf{y})|\boldsymbol{\eta}] = \nabla^2 B(\boldsymbol{\eta})$.

4.1.1 Examples of exponential family distributions

1. **Discrete distributions**

   Let's show that some of the most common distributions for random

variables that can take values on a finite or countably infinite set are part of the exponential family.

Poisson distribution

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a *Poisson distribution* let's show that $p(\mathbf{y}|\lambda)$ is in the exponential family.

$$\begin{aligned} p(\mathbf{y}|\lambda) &= \prod_{i=1}^{N} \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{N} y_i} \exp(-N\lambda)}{\prod_{i=1}^{N} y_i!} \\ &= \frac{\exp(-N\lambda) \exp(\sum_{i=1}^{N} y_i \log(\lambda))}{\prod_{i=1}^{N} y_i!}, \end{aligned}$$

then $h(\mathbf{y}) = \left[\prod_{i=1}^{N} y_i!\right]^{-1}$, $\eta(\lambda) = \log(\lambda)$, $T(\mathbf{y}) = \sum_{i=1}^{N} y_i$ (sufficient statistic) and $C(\lambda) = \exp(-\lambda)$.

If we set $\eta = \log(\lambda)$, then

$$p(\mathbf{y}|\eta) = \frac{\exp(\eta \sum_{i=1}^{N} y_i - N \exp(\eta))}{\prod_{i=1}^{N} y_i!},$$

such that $B(\eta) = N \exp(\eta)$, then $\nabla(B(\eta)) = N \exp(\eta) = N\lambda = \mathbb{E}\left[\sum_{i=1}^{N} y_i \middle| \lambda\right]$, that is, $\mathbb{E}\left[\frac{\sum_{i=1}^{N} y_i}{N} \middle| \lambda\right] = \mathbb{E}[\bar{y}|\lambda] = \lambda$, and $\nabla^2(B(\eta)) = N \exp(\eta) = N\lambda = Var\left[\sum_{i=1}^{N} y_i \middle| \lambda\right] = N^2 \times Var\left[\bar{y}|\lambda\right]$, then $Var\left[\bar{y}|\lambda\right] = \frac{\lambda}{N}$.

Bernoulli distribution

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a *Bernoulli distribution* let's show that $p(\mathbf{y}|\theta)$ is in the exponential family.

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{i=1}^{N} \theta^{y_i} (1-\theta)^{1-y_i} \\ &= \theta^{\sum_{i=1}^{N} y_i} (1-\theta)^{N - \sum_{i=1}^{N} y_i} \\ &= (1-\theta)^N \exp\left\{ \sum_{i=1}^{N} y_i \log\left(\frac{\theta}{1-\theta}\right) \right\}, \end{aligned}$$

then $h(\mathbf{y}) = \mathbb{I}[y_i \in \{0,1\}]$ (indicator function), $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$, $T(\mathbf{y}) = \sum_{i=1}^{N} y_i$ and $C(\theta) = 1-\theta$.

Write this distribution in the canonical form, and find the mean and variance of the sufficient statistic (Exercise 1).

Multinomial distribution

Given a random sample $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$ from a *m-dimensional multinomial distribution*, where $\mathbf{y}_i = [y_{i1}, \ldots, y_{im}]$, $\sum_{l=1}^{m} y_{il} = n$, n independent trials each of which leads to a success for exactly one of m categories with probabilities $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_m]$, $\sum_{l=1}^{m} \theta_l = 1$. Let's show that $p(\mathbf{y}|\boldsymbol{\theta})$ is in the exponential family.

$$
\begin{aligned}
p(\mathbf{y}|\boldsymbol{\theta}) &= \prod_{i=1}^{N} \frac{n!}{\prod_{l=1}^{m} y_{il}!} \prod_{l=1}^{m} \theta_l^{y_{il}} \\
&= \frac{(n!)^N}{\prod_{i=1}^{N}\prod_{l=1}^{m} y_{il}!} \exp\left\{ \sum_{i=1}^{N}\sum_{l=1}^{m} y_{il} \log(\theta_l) \right\} \\
&= \frac{(n!)^N}{\prod_{i=1}^{N}\prod_{l=1}^{m} y_{il}!} \exp\left\{ \left(N \times n - \sum_{i=1}^{N}\sum_{l=1}^{m-1} y_{il} \right) \log(\theta_m) \right. \\
&\left. + \sum_{i=1}^{N}\sum_{l=1}^{m-1} y_{il} \log(\theta_l) \right\} \\
&= \frac{(n!)^N}{\prod_{i=1}^{N}\prod_{l=1}^{m} y_{il}!} \theta_m^{N\times n} \exp\left\{ \sum_{i=1}^{N}\sum_{l=1}^{m-1} y_{il} \log(\theta_l/\theta_m) \right\},
\end{aligned}
$$

then $h(\mathbf{y}) = \frac{(n!)^N}{\prod_{i=1}^{N}\prod_{l=1}^{m} y_{il}!}$, $\eta(\boldsymbol{\theta}) = \left[\log\left(\frac{\theta_1}{\theta_m}\right) \ldots \log\left(\frac{\theta_{m-1}}{\theta_m}\right)\right]$, $T(\mathbf{y}) = \left[\sum_{i=1}^{N} y_{i1} \cdots \sum_{i=1}^{N} y_{im-1}\right]$ and $C(\boldsymbol{\theta}) = \theta_m^n$.

2. **Continuous distributions**

Let's show that some of the most common distributions for random variables that can take any value within a certain range or interval, often an infinite number of possible values, are part of the exponential family.

Normal distribution

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a *normal distribution* let's show that $p(\mathbf{y}|\mu, \sigma^2)$ is in the exponential family.

$$
\begin{aligned}
p(\mathbf{y}|\mu,\sigma^2) &= \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i-\mu\right)^2\right\} \\
&= (2\pi)^{-N/2}(\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-\mu\right)^2\right\} \\
&= (2\pi)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N} y_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{N} y_i \right. \\
&\left. - N\frac{\mu^2}{2\sigma^2} - \frac{N}{2}\log(\sigma^2)\right\},
\end{aligned}
$$

then $h(\mathbf{y}) = (2\pi)^{-N/2}$, $\eta(\mu,\sigma^2) = \left[\frac{\mu}{\sigma^2} \ \frac{-1}{2\sigma^2}\right]$, $T(\mathbf{y}) = \left[\sum_{i=1}^{N} y_i \ \sum_{i=1}^{N} y_i^2\right]$ and $C(\mu,\sigma^2) = \exp\left\{-\frac{\mu^2}{2\sigma^2} - \frac{\log(\sigma^2)}{2}\right\}$.
Observe that

$$
p(\mathbf{y}|\mu,\sigma^2) = (2\pi)^{-N/2} \exp\left\{\eta_1 \sum_{i=1}^{N} y_i + \eta_2 \sum_{i=1}^{N} y_i^2 - \frac{N}{2}\log(-2\eta_2) + \frac{N}{4}\frac{\eta_1^2}{\eta_2}\right\},
$$

where $B(\boldsymbol{\eta}) = \frac{N}{2}\log(-2\eta_2) - \frac{N}{4}\frac{\eta_1^2}{\eta_2}$. Then,

$$
\nabla B(\boldsymbol{\eta}) = \begin{bmatrix} -\frac{N}{2}\frac{\eta_1}{\eta_2} \\ -\frac{N}{2}\frac{1}{\eta_2} + \frac{N}{4}\frac{\eta_1^2}{\eta_2^2} \end{bmatrix} = \begin{bmatrix} N\times\mu \\ N\times(\mu^2+\sigma^2) \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[\sum_{i=1}^{N} y_i|\mu,\sigma^2\right] \\ \mathbb{E}\left[\sum_{i=1}^{N} y_i^2|\mu,\sigma^2\right] \end{bmatrix}.
$$

Multivariate normal distribution

Given $\mathbf{Y} = [\mathbf{y}_1 \ \mathbf{y}_2 \ \ldots \ \mathbf{y}_p]$ a $N \times p$ matrix such that $\mathbf{y}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, 2, \ldots, N$, that is, each i-th row of $\mathbf{Y}$ follows a *multivariate normal distribution*. Then, assuming independence between rows, let's show that $p(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is in the exponential family.

$$
\begin{aligned}
p(\mathbf{y}_1,\ldots,\mathbf{y}_N|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \prod_{i=1}^{N}(2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)\right\}\\
&= (2\pi)^{-pN/2}|\Sigma|^{-N/2}\exp\left\{-\frac{1}{2}tr\left[\sum_{i=1}^{N}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)\right]\right\}\\
&= (2\pi)^{-pN/2}|\Sigma|^{-N/2}\exp\left\{-\frac{1}{2}tr\left[\left(\mathbf{S}+N\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)^{\top}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}\\
&= (2\pi)^{-pN/2}\exp\left\{-\frac{1}{2}\left[\left(vec\left(\mathbf{S}\right)^{\top}+Nvec\left(\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\top}\right)^{\top}\right)vec\left(\boldsymbol{\Sigma}^{-1}\right)\right.\right.\\
&\quad\left.\left.-2N\hat{\boldsymbol{\mu}}^{\top}vec\left(\boldsymbol{\mu}^{\top}\boldsymbol{\Sigma}^{-1}\right)+Ntr\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\top}\boldsymbol{\Sigma}^{-1}\right)+N\log(|\boldsymbol{\Sigma}|)\right]\right\},
\end{aligned}
$$

where the second line uses the trace operator (tr), and its invariability under cyclic permutation is used in the third line. In addition, we add and subtract $\hat{\boldsymbol{\mu}}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i$ in each parenthesis such that we get $\mathbf{S}=\sum_{i=1}^{N}\left(\mathbf{y}_i-\hat{\boldsymbol{\mu}}\right)\left(\mathbf{y}_i-\hat{\boldsymbol{\mu}}\right)^{\top}$. We get the fourth line after collecting terms, and using some properties of the trace operator to introduce the vectorization operator (vec), that is, $tr(\mathbf{AB})=vec(\mathbf{A}^{\top})^{\top}vec(\mathbf{B})$, and $vec(\mathbf{A}+\mathbf{B})=vec(\mathbf{A})+vec(\mathbf{B})$.
Then $h(\mathbf{y})=(2\pi)^{-pN/2}$, $\eta(\boldsymbol{\mu},\boldsymbol{\Sigma})^{\top}=\left[\left(vec\left(\boldsymbol{\Sigma}^{-1}\right)\right)^{\top}\quad\left(vec\left(\boldsymbol{\mu}^{\top}\boldsymbol{\Sigma}^{-1}\right)\right)^{\top}\right]$, $T(\mathbf{y})=\left[-\frac{1}{2}\left(vec\left(\mathbf{S}\right)^{\top}+Nvec\left(\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\top}\right)^{\top}\right)\quad-N\hat{\boldsymbol{\mu}}^{\top}\right]^{\top}$ and $C(\boldsymbol{\mu},\boldsymbol{\Sigma})=\exp\left\{-\frac{1}{2}\left(tr\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\top}\boldsymbol{\Sigma}^{-1}\right)+\log(|\Sigma|)\right)\right\}$.

4.2 Conjugate prior to exponential family

Theorem 4.2.1

The prior distribution $\pi(\boldsymbol{\theta})\propto C(\boldsymbol{\theta})^{b_0}\exp\left\{\eta(\boldsymbol{\theta})^{\top}\mathbf{a}_0\right\}$ is conjugate to the exponential family (equation 4.4).

Proof

$$
\begin{aligned}
\pi(\boldsymbol{\theta}|\mathbf{y}) &\propto C(\boldsymbol{\theta})^{b_0}\exp\left\{\eta(\boldsymbol{\theta})^{\top}\mathbf{a}_0\right\}\times h(\mathbf{y})C(\boldsymbol{\theta})^{N}\exp\left\{\eta(\boldsymbol{\theta})^{\top}\mathbf{T}(\mathbf{y})\right\}\\
&\propto C(\boldsymbol{\theta})^{N+b_0}\exp\left\{\eta(\boldsymbol{\theta})^{\top}\left(\mathbf{T}(\mathbf{y})+\mathbf{a}_0\right)\right\}.
\end{aligned}
$$

Observe that the posterior is in the exponential family, $\pi(\boldsymbol{\theta}|\mathbf{y})\propto C(\boldsymbol{\theta})^{\beta_n}\exp\left\{\eta(\boldsymbol{\theta})^{\top}\alpha_n\right\}$, $\beta_n=N+b_0$ and $\alpha_n=\mathbf{T}(\mathbf{y})+\mathbf{a}_0$.

Remarks

We see comparing the prior and the likelihood that b_0 plays the role of a hypothetical sample size, and $\mathbf{a}_0$ plays the role of hypothetical sufficient statistics. This view helps the elicitation process.

We established the result in the *standard form* of the exponential family. We can also establish this result in the *canonical form* of the exponential family. Observe that given $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\theta})$, another way to get a prior for $\boldsymbol{\eta}$ is to use the change of variable theorem given a bijective function.

In the setting where there is a regular conjugate prior, [25] show that we obtain a posterior expectation of the sufficient statistics that is a weighted average between the prior expectation and the likelihood estimate.

4.2.1 Examples: Theorem 4.2.1

1. **Likelihood functions from discrete distributions**

   The Poisson-gamma model

   Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a Poisson distribution then a conjugate prior density for λ has the form

$$\begin{aligned}\pi(\lambda) &\propto (\exp(-\lambda))^{b_0} \exp\{a_0 \log(\lambda)\} \\ &= \exp(-\lambda b_0)\lambda^{a_0} \\ &= \exp(-\lambda\beta_0)\lambda^{\alpha_0 - 1}.\end{aligned}$$

   This is the kernel of a gamma density in the *rate parametrization*, $G(\alpha_0, \beta_0)$, $\alpha_0 = a_0 + 1$ and $\beta_0 = b_0$.[1] Then, a prior conjugate distribution for the Poisson likelihood is a gamma distribution.

   Taking into account that $\sum_{i=1}^N y_i$ is a sufficient statistic for the Poisson distribution, then we can think about a_0 as the number of occurrences in b_0 experiments. Observe that

$$\begin{aligned}\pi(\lambda|\mathbf{y}) &\propto \exp(-\lambda\beta_0)\lambda^{\alpha_0 - 1} \times \exp(-N\lambda)\lambda^{\sum_{i=1}^N y_i} \\ &= \exp(-\lambda(N + \beta_0))\lambda^{\sum_{i=1}^N y_i + \alpha_0 - 1}.\end{aligned}$$

   As expected, this is the kernel of a gamma distribution, which means $\lambda|\mathbf{y} \sim G(\alpha_n, \beta_n)$, $\alpha_n = \sum_{i=1}^N y_i + \alpha_0$ and $\beta_n = N + \beta_0$.

   Observe that α_0/β_0 is the prior mean, and α_0/β_0^2 is the prior variance. Then, $\alpha_0 \to 0$ and $\beta_0 \to 0$ imply a non-informative prior such that the posterior mean converges to the maximum likelihood estimator $\bar{y} = \frac{\sum_{i=1}^N y_i}{N}$,

[1] Another parametrization of the gamma density is the *scale parametrization* where $\kappa_0 = 1/\beta_0$. See the health insurance example in Chapter 1.

$$\begin{aligned}\mathbb{E}[\lambda|\mathbf{y}] &= \frac{\alpha_n}{\beta_n} \\ &= \frac{\sum_{i=1}^{N} y_i + \alpha_0}{N + \beta_0} \\ &= \frac{N\bar{y}}{N + \beta_0} + \frac{\alpha_0}{N + \beta_0}.\end{aligned}$$

The posterior mean is a weighted average between sample and prior information. This is a general result from regular conjugate priors [25]. Observe that $\mathbb{E}[\lambda|\mathbf{y}] = \bar{y}, \lim N \to \infty$.

In addition, $\alpha_0 \to 0$ and $\beta_0 \to 0$ corresponds to $\pi(\lambda) \propto \frac{1}{\lambda}$, which is an improper prior. Improper priors may have bad consequences on Bayes factors (hypothesis testing), see bellow a discussion of this in the linear regression framework. In this example, we can get analytical solutions for the marginal likelihood and the predictive distribution (see the health insurance example and Exercise 3 in Chapter 1).

The Bernoulli-beta model

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a Bernoulli distribution then a conjugate prior density for θ has the form

$$\begin{aligned}\pi(\theta) &\propto (1-\theta)^{b_0} \exp\left\{a_0 \log\left(\frac{\theta}{1-\theta}\right)\right\} \\ &= (1-\theta)^{b_0 - a_0}\theta^{a_0} \\ &= \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}.\end{aligned}$$

This is the kernel of a beta density, $B(\alpha_0, \beta_0)$, $\alpha_0 = a_0 + 1$ and $\beta_0 = b_0 - a_0 + 1$. A prior conjugate distribution for the Bernoulli likelihood is a beta distribution. Given that b_0 is the hypothetical sample size, and a_0 is the hypothetical sufficient statistic, which is the number of successes, then $b_0 - a_0$ is the number of failures. This implies that α_0 is the number of prior successes plus one, and β_0 is the number of prior failures plus one. Given that the mode of a beta distributed random variable is $\frac{\alpha_0 - 1}{\alpha_0 + \beta_0 - 2} = \frac{a_0}{b_0}$, then we have the prior probability of success. Setting $\alpha_0 = 1$ and $\beta_0 = 1$, which implies a 0-1 uniform distribution, corresponds to a setting with 0 successes (and 0 failures) in 0 experiments.

Observe that

$$\begin{aligned}\pi(\theta|\mathbf{y}) &\propto \theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \times \theta^{\sum_{i=1}^N y_i}(1-\theta)^{N-\sum_{i=1}^N y_i} \\ &= \theta^{\alpha_0+\sum_{i=1}^N y_i-1}(1-\theta)^{\beta_0+N-\sum_{i=1}^N y_i-1}.\end{aligned}$$

The posterior distribution is beta, $\theta|\mathbf{y} \sim B(\alpha_n, \beta_n)$, $\alpha_n = \alpha_0 + \sum_{i=1}^N y_i$ and $\beta_n = \beta_0 + N - \sum_{i=1}^N y_i$, where the posterior mean $\mathbb{E}[\theta|\mathbf{y}] = \frac{\alpha_n}{\alpha_n+\beta_n} = \frac{\alpha_0+N\bar{y}}{\alpha_0+\beta_0+N} = \frac{\alpha_0+\beta_0}{\alpha_0+\beta_0+N}\frac{\alpha_0}{\alpha_0+\beta_0} + \frac{N}{\alpha_0+\beta_0+N}\bar{y}$. The posterior mean is a weighted average between the prior mean and the maximum likelihood estimate.

El marginal likelihood in this setting is

$$\begin{aligned}p(\mathbf{y}) =& \int_0^1 \frac{\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}}{B(\alpha_0,\beta_0)} \times \theta^{\sum_{i=1}^N y_i}(1-\theta)^{N-\sum_{i=1}^N y_i} d\theta \\ =& \frac{B(\alpha_n,\beta_n)}{B(\alpha_0,\beta_0)},\end{aligned}$$

where $B(\cdot,\cdot)$ is the beta function.

In addition, the predictive density is

$$\begin{aligned}p(Y_0|\mathbf{y}) &= \int_0^1 \theta^{y_0}(1-\theta)^{1-y_0} \times \frac{\theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}}{B(\alpha_n,\beta_n)} d\theta \\ &= \frac{B(\alpha_n+y_0, \beta_n+1-y_0)}{B(\alpha_n,\beta_n)} \\ &= \frac{\Gamma(\alpha_n+\beta_n)\Gamma(\alpha_n+y_0)\Gamma(\beta_n+1-y_0)}{\Gamma(\alpha_n+\beta_n+1)\Gamma(\alpha_n)\Gamma(\beta_n)} \\ &= \left\{\begin{array}{ll}\frac{\alpha_n}{\alpha_n+\beta_n}, & y_0 = 1 \\ \frac{\beta_n}{\alpha_n+\beta_n}, & y_0 = 0\end{array}\right\}.\end{aligned}$$

This is a Bernoulli distribution with probability of success equal to $\frac{\alpha_n}{\alpha_n+\beta_n}$.

The multinomial-Dirichlet model

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a multinomial distribution then a conjugate prior density for $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_m]$ has the form

$$\begin{aligned}\pi(\boldsymbol{\theta}) &\propto \theta_m^{b_0} \exp\left\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{a}_0\right\} \\ &= \prod_{l=1}^{m-1} \theta_l^{a_{0l}} \theta_m^{b_0 - \sum_{l=1}^{m-1} a_{0l}} \\ &= \prod_{l=1}^{m} \theta_l^{\alpha_{0l}-1},\end{aligned}$$

where $\boldsymbol{\eta}(\boldsymbol{\theta}) = \left[\log\left(\frac{\theta_1}{\theta_m}\right), \ldots, \log\left(\frac{\theta_{m-1}}{\theta_m}\right)\right]$, $\mathbf{a}_0 = [a_{01}, \ldots, a_{am-1}]^\top$, $\boldsymbol{\alpha}_0 = [\alpha_{01}, \alpha_{02}, \ldots, \alpha_{0m}]$, $\alpha_{0l} = a_{0l} + 1$, $l = 1, 2, \ldots, m-1$ and $\alpha_{0m} = b_0 - \sum_{l=1}^{m-1} a_{0l} + 1$.

This is the kernel of a Dirichlet distribution, that is, the prior distribution is $D(\boldsymbol{\alpha}_0)$.

Observe that a_{0l} is the number of hypothetical number of times outcome l is observed over the hypothetical b_0 trials. Setting $\alpha_{0l} = 1$, that is a uniform distribution over the open standard simplex, implicitly we set $a_{0l} = 0$, which means that there are 0 occurrences of category l in $b_0 = 0$ experiments.

The posterior distribution of the multinomial-Dirichlet model is given by

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{y}) &\propto \prod_{l=1}^{m} \theta_l^{\alpha_{0l}-1} \times \prod_{l=1}^{m} \theta_l^{\sum_{i=1}^{N} y_{il}} \\ &= \prod_{l=1}^{m} \theta_l^{\alpha_{0l} + \sum_{i=1}^{N} y_{il} - 1}.\end{aligned}$$

This is the kernel of a Dirichlet distribution $D(\boldsymbol{\alpha}_n)$, $\boldsymbol{\alpha}_n = [\alpha_{n1}, \alpha_{n2}, \ldots, \alpha_{nm}]$, $\alpha_{nl} = \alpha_{0l} + \sum_{i=1}^{N} y_{il}$, $l = 1, 2, \ldots, m$. Observe that

$$\begin{aligned}\mathbb{E}[\theta_j|\mathbf{y}] &= \frac{\alpha_{nj}}{\sum_{l=1}^{m}\left[\alpha_{0l} + \sum_{i=1}^{N} y_{il}\right]} \\ &= \frac{\sum_{l=1}^{m} \alpha_{0l}}{\sum_{l=1}^{m}\left[\alpha_{0l} + \sum_{i=1}^{N} y_{il}\right]} \frac{\alpha_{0j}}{\sum_{l=1}^{m} \alpha_{0l}} \\ &+ \frac{\sum_{l=1}^{m}\sum_{i=1}^{N} y_{il}}{\sum_{l=1}^{m}\left[\alpha_{0l} + \sum_{i=1}^{N} y_{il}\right]} \frac{\sum_{i=1}^{N} y_{ij}}{\sum_{l=1}^{m}\sum_{i=1}^{N} y_{il}}.\end{aligned}$$

We have again that the posterior mean is a weighted average between the prior mean and the maximum likelihood estimate.

The marginal likelihood is

$$\begin{aligned} p(\mathbf{y}) &= \int_{\boldsymbol{\Theta}} \frac{\prod_{l=1}^{m} \theta_l^{\alpha_{0l}-1}}{B(\boldsymbol{\alpha}_0)} \times \prod_{i=1}^{N} \frac{n!}{\prod_{l=1}^{m} y_{il}} \prod_{l=1}^{m} \theta_l^{y_{il}} d\boldsymbol{\theta} \\ &= \frac{N \times n!}{B(\boldsymbol{\alpha}_0) \prod_{i=1}^{N} \prod_{l=1}^{m} y_{il}!} \int_{\boldsymbol{\Theta}} \prod_{l=1}^{m} \theta_l^{\alpha_{0l}+\sum_{i=1}^{N} y_{il}-1} d\boldsymbol{\theta} \\ &= \frac{N \times n!}{B(\boldsymbol{\alpha}_0) \prod_{i=1}^{N} \prod_{l=1}^{m} y_{il}!} B(\boldsymbol{\alpha}_n) \\ &= \frac{N \times n! \Gamma\left(\sum_{l=1}^{m} \alpha_{0l}\right)}{\Gamma\left(\sum_{l=1}^{m} \alpha_{0l} + N \times n\right)} \prod_{l=1}^{m} \frac{\Gamma\left(\alpha_{nl}\right)}{\Gamma\left(\alpha_{0l}\right) \prod_{i=1}^{N} y_{il}!}, \end{aligned}$$

where $B(\boldsymbol{\alpha}) = \frac{\prod_{l=1}^{m} \Gamma(\alpha_l)}{\Gamma(\sum_{l=1}^{m} \alpha_l)}$.

Following similar steps we get the predictive density

$$p(Y_0|\mathbf{y}) = \frac{n! \Gamma\left(\sum_{l=1}^{m} \alpha_{nl}\right)}{\Gamma\left(\sum_{l=1}^{m} \alpha_{nl} + n\right)} \prod_{l=1}^{m} \frac{\Gamma\left(\alpha_{nl} + y_{0l}\right)}{\Gamma\left(\alpha_{nl}\right) y_{0l}!}.$$

This is a Dirichlet-multinomial distribution with parameters $\boldsymbol{\alpha}_n$.

Example: English premier league, Liverpool vs Manchester city

Let's see an example based on data from the English Premier league. In particular, we want to get the probability that in the following five matches Liverpool versus Manchester city, the former wins two games, and the latter three game. This is done based on the historical records of the last five matches where Liverpool was local between January 14th, 2018 and April tenth, 2022. There were two wins for Liverpool, two draws, and one win for Manchester city.[2]

We use two strategies to get the hyperparameters. First, we estimate the hyperparameters of the Dirichlet distribution using betting odds from bookmakers at 19:05 hours October sixth, 2022 (Colombia time). We got information from 24 bookmakers (see file *DataOddsLIVvsMAN.csv*),[3] and transform these odds in probabilities using a simple standardization approach, then we use maximum

[2] https://www.11v11.com/teams/manchester-city/tab/opposingTeams/opposition/Liverpool/.

[3] https://www.oddsportal.com/soccer/england/premier-league/liverpool-manchester-city-WrqgEz5S/

likelihood to estimate the hyperparameters. Second, we use empirical Bayes, that is, we estimate the hyperparameters optimizing the marginal likelihood.

R code. Multinomial-Dirichlet model: Liverpool vs Manchester city

```
# Multinomial-Dirichlet example: Liverpool vs Manchester
    city
Data<-read.csv("DataApplications/DataOddsLIVvsMAN.csv", sep
    = ",", header = TRUE)
# Change path
attach(Data)
library(dplyr)
Probs <- Data %>%
  mutate(pns1 = 1/home, pns2 = 1/draw, pns3 = 1/away)%>%
  mutate(SumInvOdds = pns1 + pns2 + pns3) %>%
  mutate(p1 = pns1/SumInvOdds, p2 = pns2/SumInvOdds, p3 =
    pns3/SumInvOdds) %>%
  select(p1, p2, p3)
# We get probabilities using simple standardization. There
    are more technical approaches to do this. See for
    instance Shin (1993) and Strumbelj (2014).
DirMLE <- sirt::dirichlet.mle(Probs)
# Use maximum likelihood to estimate parameters of the
# Dirichlet distribution
alpha0odds <- DirMLE$alpha
alpha0odds
      p1       p2       p3
1599.122 1342.703 2483.129

y <- c(2, 2, 1)
# Historical records last five mathces
# Liverpool wins (2), draws (2) and Manchester
# city wins (1)

# Marginal likelihood
MarLik <- function(a0){
  n <- sum(y)
  Res1 <- sum(sapply(1:length(y),
  function(l){lgamma(a0[l]+y[l])-lgamma(a0[l])}))
  Res <- lgamma(sum(a0))-lgamma(sum(a0)+n)+Res1
  return(-Res)
}
EmpBay <- optim(alpha0odds, MarLik, method = "BFGS")
alpha0EB <- EmpBay$par
alpha0EB
p1       p2       p3
2362.622 2660.153 1279.510
# Bayes factor empirical Bayes vs betting odds.
# This is greather than 1 by construction
BF <- exp(-MarLik(alpha0EB))/exp(-MarLik(alpha0odds))
BF
2.085819
# Posterior distribution based on empirical Bayes
alphan <- alpha0EB + y
# Posterior parameters
S <- 100000
# Simulation draws from the Dirichlet distribution
thetas <- MCMCpack::rdirichlet(S, alphan)
colnames(thetas) <- c("Liverpool","Draw","Manchester")
```

R code. Multinomial-Dirichlet model: Liverpool vs Manchester city

```
# Predictive distribution based on simulations
y0 <- c(2, 0, 3)
# Liverpool two wins and Manchester city three wins in next
    five matches
Pred <- apply(thetas, 1, function(p) {rmultinom(1, size =
    sum(y0), prob = p)})
ProYo <- sum(sapply(1:S,function(s){sum(Pred[,s]==y0)==3}))/
    S
ProY0
0.0832
# Probability of y0

# Predictive distribution using analytical expression
PredY0 <- function(y0){
  n <- sum(y0)
  Res1 <- sum(sapply(1:length(y), function(l){lgamma(alphan[
    l]+y0[l]) - lgamma(alphan[l])-lfactorial(y0[l])}))
  Res <- lfactorial(n) + lgamma(sum(alphan)) - lgamma(sum(
    alphan)+n) + Res1
  return(exp(Res))
}
PredY0(y0)
0.0833
```

We see that the Bayes factor gives evidence in favor of the hyperparameters based on empirical Bayes, this is by construction, as these hyperparameters maximize the marginal likelihood.

We observe that using the hyperparameters from empirical Bayes, the probability that in the next five games Liverpool wins two games and Manchester city wins three games is 8.33%. The result using the predictive distribution based on simulations is similar to the probability using the exact predictive.

2. **Likelihood functions from continuous distributions**

   The normal-normal/inverse-gamma model

   Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a normal distribution, then the conjugate prior density has the form

$$
\begin{aligned}
\pi(\mu,\sigma^2) &\propto \exp\left\{b_0\left(-\frac{\mu^2}{2\sigma^2}-\frac{\log\sigma^2}{2}\right)\right\}\exp\left\{a_{01}\frac{\mu}{\sigma^2}-a_{02}\frac{1}{\sigma^2}\right\}\\
&= \exp\left\{b_0\left(-\frac{\mu^2}{2\sigma^2}-\frac{\log\sigma^2}{2}\right)\right\}\exp\left\{a_{01}\frac{\mu}{\sigma^2}-a_{02}\frac{1}{\sigma^2}\right\}\\
&\times \exp\left\{-\frac{a_{01}^2}{2\sigma^2 b_0}\right\}\exp\left\{\frac{a_{01}^2}{2\sigma^2 b_0}\right\}\\
&= \exp\left\{-\frac{b_0}{2\sigma^2}\left(\mu-\frac{a_{01}}{b_0}\right)^2\right\}\left(\frac{1}{\sigma^2}\right)^{\frac{b_0+1-1}{2}}\\
&\times \exp\left\{\frac{1}{\sigma^2}\frac{-2b_0a_{02}+a_{01}^2}{2b_0}\right\}\\
&= \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}}\exp\left\{-\frac{b_0}{2\sigma^2}\left(\mu-\frac{a_{01}}{b_0}\right)^2\right\}}_{1}\\
&\times \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{b_0-1}{2}}\exp\left\{-\frac{1}{\sigma^2}\frac{2b_0a_{02}-a_{01}^2}{2b_0}\right\}}_{2}.
\end{aligned}
$$

The first part is the kernel of a normal density with mean $\mu_0 = a_{01}/\beta_0$ and variance σ^2/β_0, $\beta_0 = b_0$ that is, $\mu|\sigma^2 \sim N(\mu_0, \sigma^2/\beta_0)$. The second part is the kernel of an inverse gamma density with shape parameter $\alpha_0/2 = \frac{\beta_0-3}{2}$, and scale parameter $\delta_0/2 = \frac{2\beta_0 a_{02}-a_{01}^2}{2\beta_0}$, $\sigma^2 \sim IG(\alpha_0/2, \delta_0/2)$. Observe that $b_0 = \beta_0$ is the hypothetical sample size, and a_{01} is the hypothetical sum of prior observations, then, it makes sense that a_{01}/β_0 and σ^2/β_0 are the prior mean and variance, respectively.

Therefore, the posterior distribution is also a normal-inverse gamma

distribution,

$$\begin{aligned}
\pi(\mu,\sigma^2|\mathbf{y}) &\propto \left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{\beta_0}{2\sigma^2}(\mu-\mu_0)^2\right\}\left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1}\exp\left\{-\frac{\delta_0}{2\sigma^2}\right\}\\
&\quad\times(\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right\}\\
&= \left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\beta_0(\mu-\mu_0)^2+\sum_{i=1}^{N}(y_i-\bar{y})^2+N(\mu-\bar{y})^2+\delta_0\right)\right\}\\
&\quad\times\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0+N}{2}+1}+\frac{(\beta_0\mu_0+N\bar{y})^2}{\beta_0+N}-\frac{(\beta_0\mu_0+N\bar{y})^2}{\beta_0+N}\\
&= \underbrace{\left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left((\beta_0+N)\left(\mu-\left(\frac{\beta_0\mu_0+N\bar{y}}{\beta_0+N}\right)\right)^2\right)\right\}}_{1}\\
&\quad\times\underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0+N}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}(y_i-\bar{y})^2+\delta_0+\frac{\beta_0 N}{\beta_0+N}(\bar{y}-\mu_0)^2\right)\right\}}_{2}.
\end{aligned}$$

The first term is the kernel of a normal density, $\mu|\sigma^2,\mathbf{y}\sim N\left(\mu_n,\sigma_n^2\right)$, where $\mu_n=\frac{\beta_0\mu_0+N\bar{y}}{\beta_0+N}$ and $\sigma_n^2=\frac{\sigma^2}{\beta_n}$, $\beta_n=\beta_0+N$. The second term is the kernel of an inverse gamma density, $\sigma^2|\mathbf{y}\sim IG(\alpha_n/2,\delta_n/2)$ where $\alpha_n=\alpha_0+N$ and $\delta_n=\sum_{i=1}^{N}(y_i-\bar{y})^2+\delta_0+\frac{\beta_0 N}{\beta_0+N}(\bar{y}-\mu_0)^2$. Observe that the posterior mean is a weighted average between prior and sample information. The weights depends on the sample sizes (β_0 and N).

The marginal posterior for σ^2 is inverse gamma with shape and scale parameters $\alpha_n/2$ and $\delta_n/2$, respectively. The marginal posterior of μ is

$$\begin{aligned}
\pi(\mu|\mathbf{y}) &\propto \int_0^\infty\left\{\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_n+1}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}(\beta_n(\mu-\mu_n)^2+\delta_n)\right\}\right\}d\sigma^2\\
&= \frac{\Gamma\left(\frac{\alpha_n+1}{2}\right)}{\left[\frac{\beta_n(\mu-\mu_n)^2+\delta_n}{2}\right]^{\frac{\alpha_n+1}{2}}}\\
&\propto \left[\frac{\beta_n(\mu-\mu_n)^2+\delta_n}{2}\right]^{-\frac{\alpha_n+1}{2}}\left(\frac{\delta_n}{\delta_n}\right)^{-\frac{\alpha_n+1}{2}}\\
&\propto \left[\frac{\alpha_n\beta_n(\mu-\mu_n)^2}{\alpha_n\delta_n}+1\right]^{-\frac{\alpha_n+1}{2}},
\end{aligned}$$

where the second line due to having the kernel of an inverse gamma density with parameters $(\alpha_n+1)/2$ and $-\frac{1}{2\sigma^2}(\beta_n(\mu-\mu_n)^2+\delta_n)$.

This is the kernel of a Student's t distribution, $\mu|\mathbf{y} \sim t(\mu_n, \delta_n/\beta_n\alpha_n, \alpha_n)$, where $\mathbb{E}[\mu|\mathbf{y}] = \mu_n$ and $Var[\mu|\mathbf{y}] = \frac{\alpha_n}{\alpha_n-2}\left(\frac{\delta_n}{\beta_n\alpha_n}\right) = \frac{\delta_n}{(\alpha_n-2)\beta_n}$, $\alpha_n > 2$. Observe that the marginal posterior distribution for μ has heavier tails than the conditional posterior distribution due to incorporating uncertainty regarding σ^2. The marginal likelihood is

$$
\begin{aligned}
p(\mathbf{y}) &= \int_{-\infty}^{\infty}\int_{0}^{\infty}\left\{(2\pi\sigma^2/\beta_0)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2/\beta_0}(\mu-\mu_0)^2\right\}\frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}\left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1}\right. \\
&\quad \left.\times\exp\left\{-\frac{\delta_0}{2\sigma^2}\right\}(2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right\}\right\}d\sigma^2 d\mu \\
&= \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}(2\pi)^{-\left(\frac{N+1}{2}\right)}\beta_0^{1/2}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left\{\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0+N+1}{2}+1}\right. \\
&\quad \left.\times\exp\left\{-\frac{1}{2\sigma^2}(\beta_0(\mu-\mu_0)^2+\sum_{i=1}^{N}(y_i-\mu)^2+\delta_0)\right\}\right\}d\sigma^2 d\mu \\
&= \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}(2\pi)^{-\left(\frac{N+1}{2}\right)}\beta_0^{1/2}\Gamma\left(\frac{N+1+\alpha_0}{2}\right) \\
&\quad \times\int_{-\infty}^{\infty}\left[\frac{\beta_0(\mu-\mu_0)^2+\sum_{i=1}^{N}(y_i-\mu)^2+\delta_0}{2}\right]^{-\frac{\alpha_0+N+1}{2}}d\mu \\
&= \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}(2\pi)^{-\left(\frac{N+1}{2}\right)}\beta_0^{1/2}\Gamma\left(\frac{N+1+\alpha_0}{2}\right) \\
&\quad \times\int_{-\infty}^{\infty}\left[\frac{\beta_n(\mu-\mu_n)^2+\delta_n}{2}\right]^{-\frac{\alpha_n+1}{2}}d\mu\left(\frac{\delta_n/2}{\delta_n/2}\right)^{-\frac{\alpha_n+1}{2}} \\
&= \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}(2\pi)^{-\left(\frac{N+1}{2}\right)}\beta_0^{1/2}\Gamma\left(\frac{\alpha_n+1}{2}\right)\left(\frac{\delta_n}{2}\right)^{-\frac{\alpha_n+1}{2}}\frac{\left(\frac{\delta_n\pi}{\beta_n}\right)^{1/2}\Gamma\left(\frac{\alpha_n}{2}\right)}{\Gamma\left(\frac{\alpha_n+1}{2}\right)} \\
&= \frac{\Gamma\left(\frac{\alpha_n}{2}\right)}{\Gamma\left(\frac{\alpha_0}{2}\right)}\frac{(\delta_0/2)^{\alpha_0/2}}{(\delta_n/2)^{\alpha_n/2}}\left(\frac{\beta_0}{\beta_n}\right)^{1/2}(\pi)^{-N/2},
\end{aligned}
$$

where we take into account that $\int_{-\infty}^{\infty}\left[\frac{\beta_n(\mu-\mu_n)^2+\delta_n}{2}\right]^{-\frac{\alpha_n+1}{2}}d\mu\left(\frac{\delta_n/2}{\delta_n/2}\right)^{-\frac{\alpha_n+1}{2}} = \int_{-\infty}^{\infty}\left[\frac{\beta_n\alpha_n(\mu-\mu_n)^2}{\delta_n\alpha_n}+1\right]^{-\frac{\alpha_n+1}{2}}d\mu\left(\frac{\delta_n}{2}\right)^{-\frac{\alpha_n+1}{2}}$. The term in the integral is the kernel of a Student's t density, this means that the integral is equal to $\frac{\left(\frac{\delta_n\pi}{\beta_n}\right)^{1/2}\Gamma\left(\frac{\alpha_n}{2}\right)}{\Gamma\left(\frac{\alpha_n+1}{2}\right)}$.

The predictive density is

$$\begin{aligned}
\pi(Y_0|\mathbf{y}) &\propto \int_{-\infty}^{\infty}\int_{0}^{\infty}\left\{\left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(y_0-\mu)^2\right\}\left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{\beta_n}{2\sigma^2}(\mu-\mu_n)^2\right\}\right.\\
&\left.\times\left(\frac{1}{\sigma^2}\right)^{\alpha_n/2+1}\exp\left\{-\frac{\delta_n}{2\sigma^2}\right\}\right\}d\sigma^2 d\mu\\
&=\int_{-\infty}^{\infty}\int_{0}^{\infty}\left\{\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_n+2}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}((y_0-\mu)^2+\beta_n(\mu-\mu_n)^2+\delta_n)\right\}\right\}d\sigma^2 d\mu\\
&\propto\int_{-\infty}^{\infty}\left[\beta_n(\mu-\mu_n)^2+(y_0-\mu)^2+\delta_n\right]^{-\left(\frac{\alpha_n}{2}+1\right)}d\mu\\
&=\int_{-\infty}^{\infty}\left[(\beta_n+1)\left(\mu-\left(\frac{\beta_n\mu_n+y_0}{\beta_n+1}\right)\right)^2+\frac{\beta_n(y_0-\mu_n)^2}{\beta_n+1}+\delta_n\right]^{-\left(\frac{\alpha_n}{2}+1\right)}d\mu\\
&=\int_{-\infty}^{\infty}\left[1+\frac{(\beta_n+1)^2\left(\mu-\left(\frac{\beta_n\mu_n+y_0}{\beta_n+1}\right)\right)^2}{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}\right]^{-\left(\frac{\alpha_n}{2}+1\right)}d\mu\\
&\times\left(\frac{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}{\beta_n+1}\right)^{-\left(\frac{\alpha_n}{2}+1\right)}\\
&\propto\left(\frac{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}{(\beta_n+1)^2(\alpha_n+1)}\right)^{\frac{1}{2}}\left(\frac{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}{\beta_n+1}\right)^{-\left(\frac{\alpha_n}{2}+1\right)}\\
&\propto(\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n)^{\left(\frac{\alpha_n+1}{2}\right)}\\
&\propto\left[1+\frac{\beta_n\alpha_n}{(\beta_n+1)\delta_n\alpha_n}(y_0-\mu_n)^2\right]^{-\left(\frac{\alpha_n+1}{2}\right)},
\end{aligned}$$

where we have that $\left[1+\frac{(\beta_n+1)^2\left(\mu-\left(\frac{\beta_n\mu_n+y_0}{\beta_n+1}\right)\right)^2}{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}\right]^{-\left(\frac{\alpha_n}{2}+1\right)}$ is the kernel of a Student's t density with degrees of freedom α_n+1 and scale $\frac{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}{(\beta_n+1)^2(\alpha_n+1)}$.

The last expression is the kernel of a Student's t density, that is, $Y_0|\mathbf{y}\sim t\left(\mu_n,\frac{(\beta_n+1)\delta_n}{\beta_n\alpha_n},\alpha_n\right)$.

The multivariate normal-normal/inverse-Wishart model

We show in subsection 4.1 that the multivariate normal distribution is in the exponential family where

$$C(\boldsymbol{\mu},\boldsymbol{\Sigma})=\exp\left\{-\frac{1}{2}\left(tr\left(\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\right)+\log(|\Sigma|)\right)\right\},$$

$$\eta(\boldsymbol{\mu},\boldsymbol{\Sigma})^\top=\left[\left(vec\left(\boldsymbol{\Sigma}^{-1}\right)\right)^\top\quad\left(vec\left(\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\right)\right)^\top\right],$$

$$T(\mathbf{y})=\left[-\frac{1}{2}\left(vec\left(\mathbf{S}\right)^\top+Nvec\left(\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top\right)^\top\right)\quad-N\hat{\boldsymbol{\mu}}^\top\right]^\top$$

and

$$h(\mathbf{y}) = (2\pi)^{-pN/2}.$$

Then, its conjugate prior distribution should have the form

$$\begin{aligned}
\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto \exp\left\{-\frac{b_0}{2}\left(tr\left(\boldsymbol{\mu}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\right) + \log(|\Sigma|)\right)\right\} \\
&\times \exp\left\{\mathbf{a}_{01}^\top vec\left(\boldsymbol{\Sigma}^{-1}\right) + \mathbf{a}_{02}^\top vec\left(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\right)\right\} \\
&= |\Sigma|^{-b_0/2} \exp\left\{-\frac{b_0}{2}\left(tr\left(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right) + tr\left(\mathbf{a}_{02}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\} \\
&\times \exp\left\{\mathbf{a}_{01}^\top vec\left(\boldsymbol{\Sigma}^{-1}\right) + \frac{\mathbf{a}_{02}^\top \boldsymbol{\Sigma}^{-1}\mathbf{a}_{02}}{2b_0} - \frac{\mathbf{a}_{02}^\top \boldsymbol{\Sigma}^{-1}\mathbf{a}_{02}}{2b_0}\right\} \\
&= |\Sigma|^{-b_0/2} \exp\left\{-\frac{b_0}{2}\left(\boldsymbol{\mu} - \frac{\mathbf{a}_{02}}{b_0}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} - \frac{\mathbf{a}_{02}}{b_0}\right)\right\} \\
&\times \exp\left\{-\frac{1}{2} tr\left(\left(\mathbf{A}_{01} - \frac{\mathbf{a}_{02}\mathbf{a}_{02}^\top}{b_0}\right)\boldsymbol{\Sigma}^{-1}\right)\right\} \\
&= \underbrace{|\Sigma|^{-1/2} \exp\left\{-\frac{b_0}{2}\left(\boldsymbol{\mu} - \frac{\mathbf{a}_{02}}{b_0}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} - \frac{\mathbf{a}_{02}}{b_0}\right)\right\}}_{1} \\
&\times \underbrace{|\Sigma|^{-(\alpha_0+p+1)/2} \exp\left\{-\frac{1}{2} tr\left(\left(\mathbf{A}_{01} - \frac{\mathbf{a}_{02}\mathbf{a}_{02}^\top}{b_0}\right)\boldsymbol{\Sigma}^{-1}\right)\right\}}_{2},
\end{aligned}$$

where b_0 is the hypothetical sample size, and $\mathbf{a}_{01}$ and $\mathbf{a}_{02}$ are p^2 and p dimensional vectors of prior sufficient statistics, where $\mathbf{a}_{01} = -\frac{1}{2}vec(\mathbf{A}_{01})$ such that $\mathbf{A}_{01}$ is a $p \times p$ positive semi-definite matrix. Setting $b_0 = 1+\alpha_0+p+1$, we have that the first part in the last expression is the kernel of a multivariate normal density with mean $\boldsymbol{\mu}_0 = \mathbf{a}_{02}/b_0$ and covariance $\frac{\boldsymbol{\Sigma}}{b_0}$, that is, $\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim N_p\left(\boldsymbol{\mu}_0, \frac{\boldsymbol{\Sigma}}{\beta_0}\right)$, $b_0 = \beta_0$. It makes sense these hyperparameters because $\mathbf{a}_{02}$ is the hypothetical sum of prior observations and b_0 is the hypothetical prior sample size. In addition, the second expression in the last line is the kernel of a inverse Wishart distribution with scale matrix $\boldsymbol{\Psi}_0 = \left(\mathbf{A}_{01} - \frac{\mathbf{a}_{02}\mathbf{a}_{02}^\top}{b_0}\right)$ and α_0 degrees of freedom, that is, $\boldsymbol{\Sigma} \sim IW_p(\boldsymbol{\Psi}_0, \alpha_0)$. Observe that $\boldsymbol{\Psi}_0$ has the same structure as the first part of the sufficient statistics in $T(\mathbf{y})$, just that it should be understood as coming from prior hypothetical observations.

Therefore, the prior distribution in this setting is normal/inverse-Wishart, and given conjugacy, the posterior distribution is in the

same family.

$$\pi(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{Y}) \propto (2\pi)^{-pN/2}|\Sigma|^{-N/2}\exp\left\{-\frac{1}{2}tr\left[\left(\mathbf{S}+N\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)^{\top}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}$$
$$\times|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{\beta_0}{2}tr\left[(\boldsymbol{\mu}-\boldsymbol{\mu}_0)(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\top}\boldsymbol{\Sigma}^{-1}\right]\right\}|\boldsymbol{\Sigma}|^{-(\alpha_0+p+1)/2}$$
$$\times\exp\left\{-\frac{1}{2}tr(\boldsymbol{\Psi}_0\boldsymbol{\Sigma}^{-1})\right\}.$$

Taking into account that

$$N\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)^{\top}+\beta_0\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)^{\top}=\left(N+\beta_0\right)\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)^{\top}$$
$$+\frac{N\beta_0}{N+\beta_0}\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\right)\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\right)^{\top},$$

where $\boldsymbol{\mu}_n=\frac{N}{N+\beta_0}\hat{\boldsymbol{\mu}}+\frac{\beta_0}{N+\beta_0}\boldsymbol{\mu}_0$ is the posterior mean. We have

$$\pi(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{Y}) \propto |\Sigma|^{-1/2}\exp\left\{-\frac{N+\beta_0}{2}tr\left[\left(\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)^{\top}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}$$
$$\times|\boldsymbol{\Sigma}|^{-(N+\alpha_0+p+1)/2}$$
$$\times\exp\left\{-\frac{1}{2}tr\left[\left(\boldsymbol{\Psi}_0+\mathbf{S}+\frac{N\beta_0}{N+\beta_0}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0)(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0)^{\top}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}.$$

Then, $\boldsymbol{\mu}|\boldsymbol{\Sigma},\mathbf{Y}\sim N_p\left(\boldsymbol{\mu}_n,\frac{1}{\beta_n}\boldsymbol{\Sigma}\right)$, and $\boldsymbol{\Sigma}|\mathbf{Y}\sim IW\left(\boldsymbol{\Psi}_n,\alpha_n\right)$ where $\beta_n=N+\beta_0$, $\alpha_n=N+\alpha_0$ and $\boldsymbol{\Psi}_n=\boldsymbol{\Psi}_0+\mathbf{S}+\frac{N\beta_0}{N+\beta_0}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0)(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0)^{\top}$.

The marginal posterior of $\boldsymbol{\mu}$ is given by $\int_{\mathcal{S}}\pi(\boldsymbol{\mu},\boldsymbol{\Sigma})d\boldsymbol{\Sigma}$ where $\mathcal{S}$ is the space of positive semi-definite matrices. Then,

$$\pi(\boldsymbol{\mu}|\mathbf{Y}) \propto \int_{\mathcal{S}}\left\{|\boldsymbol{\Sigma}|^{-(\alpha_n+p+2)/2}\right.$$
$$\left.\exp\left\{-\frac{1}{2}tr\left[\left(\beta_n\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)^{\top}+\boldsymbol{\Psi}_n\right)\boldsymbol{\Sigma}^{-1}\right]\right\}\right\}d\boldsymbol{\Sigma}$$
$$\propto\left|\left(\beta_n\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)^{\top}+\boldsymbol{\Psi}_n\right)\right|^{-(\alpha_n+1)/2}$$
$$=\left[\left|\boldsymbol{\Psi}_n\right|\times\left|1+\beta_n\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)^{\top}\boldsymbol{\Psi}_n^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)\right|\right]^{-(\alpha_n+1)/2}$$
$$\propto\left(1+\frac{1}{\alpha_n+1-p}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)^{\top}\left(\frac{\boldsymbol{\Psi}_n}{(\alpha_n+1-p)\beta_n}\right)^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)\right)^{-(\alpha_n+1-p+p)/2},$$

where the second line uses properties of the inverse Wishart distribution, and the third line uses a particular case of the Sylvester's determinant theorem.

We observe that the last line is the kernel of a multivariate t distribution, that is, $\boldsymbol{\mu}|\mathbf{Y}\sim t_p(v_n,\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)$ where $v_n=\alpha_n+1-p$ and $\boldsymbol{\Sigma}_n=\frac{\boldsymbol{\Psi}_n}{(\alpha_n+1-p)\beta_n}$.

The marginal likelihood is given by

$$p(\mathbf{Y})=\frac{\Gamma_p\left(\frac{v_n}{2}\right)}{\Gamma_p\left(\frac{\alpha_0}{2}\right)}\frac{|\boldsymbol{\Psi}_0|^{\alpha_0/2}}{|\boldsymbol{\Psi}_n|^{\alpha_n/2}}\left(\frac{\beta_0}{\beta_n}\right)^{p/2}(2\pi)^{-Np/2},$$

where Γ_p is the multivariate gamma function (see Exercise 5).

The posterior predictive distribution is $\mathbf{Y}_0|\mathbf{Y} \sim t_p(v_n, \boldsymbol{\mu}_n, (\beta_n + 1)\boldsymbol{\Sigma}_n)$ (see Exercise 6).

Example: Tangency portfolio of US tech stocks

The tangency portfolio is the portfolio that maximizes the Sharpe ratio, where this is the excess of return of a portfolio standardized by its risk.

We want to find the shares $\mathbf{w}$ of a portfolio that maximizes the Sharpe ratio, where $\mu_{i,T+\kappa} = \mathbb{E}(R_{i,T+\kappa} - R_{f,T+\kappa} \mid \mathcal{I}_T)$, $R_{i,T+\kappa}$ and $R_{f,T+\kappa}$ are the returns of stock i and a risk-free asset. Observe that we have the expected value at period $T+\kappa$ of the excess return conditional on information up to T ($\mathcal{I}_T$), and $\boldsymbol{\Sigma}_{T+\kappa}$ is the covariance of the excess returns, which is a measure of risk. In particular,

$$\arg\max_{\mathbf{w}\in\mathbb{R}^p} \frac{\mathbf{w}^\top \boldsymbol{\mu}_{T+\kappa}}{\sqrt{\mathbf{w}^\top \boldsymbol{\Sigma}_{T+\kappa}\mathbf{w}}}; \qquad \text{s.t} \quad \mathbf{w}^\top \mathbf{1} = 1,$$

where the solution is

$$\mathbf{w}^* = \frac{\boldsymbol{\Sigma}_{T+\kappa}^{-1}\boldsymbol{\mu}_{T+\kappa}}{\mathbf{1}^\top \boldsymbol{\Sigma}_{T+\kappa}^{-1}\boldsymbol{\mu}_{T+\kappa}}.$$

If we want to find the optimal portfolio for the next period under the assumption that the excess of returns follow a multivariate normal distribution, which is a common assumption in these applications, we can set $\kappa = 1$, and use the predictive distribution of the excess of returns such that $\boldsymbol{\mu}_{T+1} = \boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_{T+1} = \frac{v_n}{v_n-2}(\beta_n+1)\boldsymbol{\Sigma}_n$ given the previous predictive result.

We apply this framework to ten tech stocks of the US market between January first, 2021, and September ninth, 2022. In particular, we use information from Yahoo Finance for Apple (AAPL), Netflix (NFLX), Amazon (AMZN), Microsoft (MSFT), Google (GOOG), Meta (META), Tesla (TSLA), NVIDIA Corporation (NVDA), Intel (INTC), and PayPal (PYPL).

R code. Optimal tangency portfolio: Tech shares

```
library(quantmod)
library(xts)
library(ggplot2)
library(gridExtra)
# grid.arrange
graphics.off()
rm(list=ls())
# Data Range
sdate <- as.Date("2021-01-01")
edate <- as.Date("2022-09-30")
Date <- seq(sdate, edate, by = "day")
tickers <- c("AAPL", "NFLX", "AMZN", "GOOG", "INTC","META",
    "MSFT", "TSLA", "NVDA", "PYPL")
p <- length(tickers)
# AAPL: Apple, NFLX: Netflix, AMZN: Amazon,
# MSFT: Microsoft, GOOG: Google, META: Meta,
# TSLA: Tesla, NVDA: NVIDIA Corporation
# INTC: Intel, PYPL: PayPal
ss_stock <- getSymbols(tickers, from=sdate, to=edate, auto.
    assign = T)
ss_stock <- purrr::map(tickers,function(x) Ad(get(x)))
ss_stock <- as.data.frame(purrr::reduce(ss_stock, merge))
colnames(ss_stock) <- tickers
# This is to get stock prices
ss_rtn <- as.data.frame(apply(ss_stock, 2, function(x) {diff
    (log(x), 1)}))
# Daily returns
t10yr <- getSymbols(Symbols = "DGS10", src = "FRED", from=
    sdate, to=edate, auto.assign = F)
# To get 10-Year US Treasury yield data from the
Federal Reserve Electronic Database (FRED)
t10yrd <- (1 + t10yr/100)^(1/365)-1
# Daily returns
t10yrd <- t10yrd[row.names(ss_rtn)]
Exc_rtn <- as.matrix(ss_rtn) - kronecker(t(rep(1, p)), as.
    matrix(t10yrd))
# Excesses of return
df <- as.data.frame(Exc_rtn)
df$Date <- as.Date(rownames(df))
#  Get months
df$Month <- months(df$Date)
#  Get years
df$Year <- format(df$Date, format="%y")
#  Aggregate on months and year and get mean
Data <- sapply(1:p, function(i) {
  aggregate(df[, i] ~ Month + Year, df, mean)})
DataExcRtn <- matrix(0, length(Data[, 1]$Month), p)
for(i in 1:p){
  DataExcRtn[, i] <- as.numeric(Data[, i]$`df[, i]`)
}
colnames(DataExcRtn) <- tickers
head(DataExcRtn)
```

R code. Optimal tangency portfolio: Tech shares

```
# Hyperparameters #
N <- dim(DataExcRtn)[1]
mu0 <- rep(0, p)
beta0 <- 1
Psi0 <- 100 * diag(p)
alpha0 <- p + 2
# Posterior parameters #
alphan <- N + alpha0
vn <- alphan + 1 - p
muhat <- colMeans(DataExcRtn)
mun <- N/(N + beta0) * muhat + beta0/(N + beta0) * mu0
S <- t(DataExcRtn - rep(1, N)%*%t(muhat))%*%(DataExcRtn -
    rep(1, N) %*%t(muhat))
Psin <- Psi0 + S + N*beta0/(N + beta0)*(muhat - mu0)%*%t(
    muhat - mu0)
betan <- N + beta0
Sigman <- Psin/((alphan + 1 - p)*betan)
Covarn <- (Sigman * (1 + betan)) * vn / (vn - 2)
Covari <- solve(Covarn)
OptShare <- t(Covari%*%mun/as.numeric((t(rep(1, p))%*%Covari
    %*%mun)))
colnames(OptShare) <- tickers
OptShare
  AAPL  NFLX  AMZN   GOOG INTC  META MSFT    TSLA  NVDA
    PYPL
-0.019 0.248 0.102 -0.034 0.173 0.23 -0.022 -0.016 0.035
    0.301
```

We find that the optimal tangency portfolio is composed by 24.8%, 10.2%, 17.3%, 23%, 3.5% and 30.1% weights of Netflix, Amazon, Intel, Meta, NVIDIA and PayPal, and -1.9%, -3.4%, -2.2% and -1.6% weights of Apple, Google, Microsoft and Tesla. A negative weight means being short in financial jargon, that is, borrowing a stock to sell it.

4.3 Linear regression: The conjugate normal-normal/inverse gamma model

In this setting we analyze the conjugate normal-normal/inverse gamma model which is the workhorse in econometrics. In this model, the dependent variable

y_i is related to a set of regressors $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})^\top$ in a linear way, that is, $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \mu_i = \mathbf{x}_i^\top \beta + \mu_i$ where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)^\top$ and $\mu_i \overset{iid}{\sim} N(0, \sigma^2)$ is an stochastic error such that $\mathbb{E}[\mu_i | \mathbf{x}_i] = 0$.

Defining $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1K} \\ x_{21} & x_{22} & \ldots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{NK} \end{bmatrix}$ and $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}$,

we can write the model in matrix form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mu$, where $\mu \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. This implies that $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$. Then, the likelihood function is

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\ &\propto (\sigma^2)^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}. \end{aligned}$$

The conjugate priors for the parameters are

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2 &\sim N(\boldsymbol{\beta}_0, \sigma^2 \mathbf{B}_0), \\ \sigma^2 &\sim IG(\alpha_0/2, \delta_0/2). \end{aligned}$$

Then, the posterior distribution is

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto (\sigma^2)^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\ &\quad \times (\sigma^2)^{-\frac{K}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{B}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\} \\ &\quad \times \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)} \left(\frac{1}{\sigma^2} \right)^{\alpha_0/2+1} \exp\left\{ -\frac{\delta_0}{2\sigma^2} \right\} \\ &\propto (\sigma^2)^{-\frac{K}{2}} \exp\left\{ -\frac{1}{2\sigma^2} [\boldsymbol{\beta}^\top (\mathbf{B}_0^{-1} + \mathbf{X}^\top \mathbf{X}) \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top (\mathbf{B}_0^{-1} \boldsymbol{\beta}_0 + \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}})] \right\} \\ &\quad \times \left(\frac{1}{\sigma^2} \right)^{(\alpha_0+N)/2+1} \exp\left\{ -\frac{\delta_0 + \mathbf{y}^\top \mathbf{y} + \boldsymbol{\beta}_0^\top \mathbf{B}_0^{-1} \boldsymbol{\beta}_0}{2\sigma^2} \right\}, \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ is the maximum likelihood estimator.

Adding and subtracting $\boldsymbol{\beta}_n^\top \mathbf{B}_n^{-1} \boldsymbol{\beta}_n$ to complete the square, where $\mathbf{B}_n = (\mathbf{B}_0^{-1} + \mathbf{X}^\top \mathbf{X})^{-1}$ and $\boldsymbol{\beta}_n = \mathbf{B}_n (\mathbf{B}_0^{-1} \boldsymbol{\beta}_0 + \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}})$,

$$\pi(\boldsymbol{\beta},\sigma^2|\mathbf{y},\mathbf{X}) \propto \underbrace{(\sigma^2)^{-\frac{K}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\top \mathbf{B}_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\right\}}_{1}$$
$$\times \underbrace{(\sigma^2)^{-\left(\frac{\alpha_n}{2}+1\right)} \exp\left\{-\frac{\delta_n}{2\sigma^2}\right\}}_{2}.$$

The first expression is the kernel of a normal density function, $\boldsymbol{\beta}|\sigma^2,\mathbf{y},\mathbf{X} \sim N(\boldsymbol{\beta}_n,\sigma^2\mathbf{B}_n)$. The second expression is the kernel of a inverse gamma density, $\sigma^2|\mathbf{y},\mathbf{X} \sim IG(\alpha_n/2,\delta_n/2)$, where $\alpha_n = \alpha_0 + N$ and $\delta_n = \delta_0 + \mathbf{y}^\top\mathbf{y} + \boldsymbol{\beta}_0^\top\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_n^\top\mathbf{B}_n^{-1}\boldsymbol{\beta}_n$.

Taking into account that

$$\begin{aligned}\boldsymbol{\beta}_n &= (\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0+\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}})\\ &= (\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{B}_0^{-1}\boldsymbol{\beta}_0+(\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}},\end{aligned}$$

where $(\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{B}_0^{-1} = \mathbf{I_K} - (\mathbf{B_0^{-1}}+\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X}$ [84]. Setting $\mathbf{W} = (\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}$ we have $\boldsymbol{\beta}_n = (\mathbf{I_K}-\mathbf{W})\boldsymbol{\beta_0} + \mathbf{W}\hat{\boldsymbol{\beta}}$, that is, the posterior mean of $\boldsymbol{\beta}$ is a weighted average between the sample and prior information, where the weights depend on the precision of each piece of information. Observe that when the prior covariance matrix is highly vague (non–informative), such that $\mathbf{B}_0^{-1} \to \mathbf{0_K}$, we obtain $\mathbf{W} \to I_K$, such that $\boldsymbol{\beta}_n \to \hat{\boldsymbol{\beta}}$, that is, the posterior mean location parameter converges to the maximum likelihood estimator.

In addition, we know that the posterior conditional covariance matrix of the location parameters $\sigma^2(\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} - \sigma^2\left((\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{B}_0+(\mathbf{X}^\top\mathbf{X})^{-1})^{-1}(\mathbf{X}^\top\mathbf{X})^{-1}\right)$ is positive semi-definite.[4] Given that $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ is the covariance matrix of the maximum likelihood estimator, we observe that prior information reduces estimation uncertainty.

Now, we calculate the posterior marginal distribution of $\boldsymbol{\beta}$,

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) &= \int_0^\infty \pi(\boldsymbol{\beta},\sigma^2|\mathbf{y},\mathbf{X})d\sigma^2\\ &= \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_n+K}{2}+1} \exp\left\{-\frac{s}{2\sigma^2}\right\}d\sigma^2,\end{aligned}$$

where $s = \delta_n + (\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\top\mathbf{B}_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)$. Then we can write

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) &= \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_n+K}{2}+1} \exp\left\{-\frac{s}{2\sigma^2}\right\}d\sigma^2\\ &= \frac{\Gamma((\alpha_n+K)/2)}{(s/2)^{(\alpha_n+K)/2}}\int_0^\infty \frac{(s/2)^{(\alpha_n+K)/2}}{\Gamma((\alpha_n+K)/2)}(\sigma^2)^{-(\alpha_n+K)/2-1}\exp\left\{-\frac{s}{2\sigma^2}\right\}d\sigma^2.\end{aligned}$$

[4] A particular case of the Woodbury matrix identity.

The right term is the integral of the probability density function of an inverse gamma distribution with parameters $\nu = (\alpha_n + K)/2$ and $\tau = s/2$. Since we are integrating over the whole support of σ^2, the integral is equal to 1, and therefore

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) &= \frac{\Gamma((\alpha_n + K)/2)}{(s/2)^{(\alpha_n+K)/2}} \\
&\propto s^{-(\alpha_n+K)/2} \\
&= [\delta_n + (\boldsymbol{\beta} - \boldsymbol{\beta}_n)^\top \mathbf{B}_n^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)]^{-(\alpha_n+K)/2} \\
&= \left[1 + \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}_n)^\top \left(\frac{\delta_n}{\alpha_n}\mathbf{B}_n\right)^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)}{\alpha_n}\right]^{-(\alpha_n+K)/2} (\delta_n)^{-(\alpha_N+K)/2} \\
&\propto \left[1 + \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}_n)^\top \mathbf{H}_n^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)}{\alpha_n}\right]^{-(\alpha_n+K)/2},
\end{aligned}$$

where $\mathbf{H}_n = \frac{\delta_n}{\alpha_n}\mathbf{B}_n$. This last expression is a multivariate t distribution for $\boldsymbol{\beta}$, $\boldsymbol{\beta}|\mathbf{y}, \mathbf{X} \sim t_K(\alpha_n, \boldsymbol{\beta}_n, \mathbf{H}_n)$.

Observe that as we have incorporated the uncertainty of the variance, the posterior for $\boldsymbol{\beta}$ changes from a normal to a t distribution, which has heavier tails, indicating more uncertainty.

The marginal likelihood of this model is

$$p(\mathbf{y}) = \int_0^\infty \int_{R^K} \pi(\boldsymbol{\beta}|\sigma^2, \mathbf{B}_0, \boldsymbol{\beta}_0)\pi(\sigma^2|\alpha_0/2, \delta_0/2)p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X})d\sigma^2 d\boldsymbol{\beta}.$$

Taking into account that $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{B}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = (\boldsymbol{\beta} - \boldsymbol{\beta}_n)^\top \mathbf{B}_n^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_n) + m$, where $m = \mathbf{y}^\top\mathbf{y} + \boldsymbol{\beta}_0^\top \mathbf{B}_0^{-1}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_n^\top \mathbf{B}_n^{-1}\boldsymbol{\beta}_n$, we have that

$$\begin{aligned}
p(\mathbf{y}) &= \int_0^\infty \int_{R^K} \pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2)p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X})d\sigma^2 d\boldsymbol{\beta} \\
&= \int_0^\infty \pi(\sigma^2)\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}m\right\}\frac{1}{(2\pi\sigma^2)^{K/2}|\mathbf{B}_0|^{1/2}} \\
&\times \int_{R^K} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)^\top \mathbf{B}_n^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)\right\} d\sigma^2 d\boldsymbol{\beta} \\
&= \int_0^\infty \pi(\sigma^2)\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}m\right\}\frac{|\mathbf{B}_n|^{1/2}}{|\mathbf{B}_0|^{1/2}}d\sigma^2 \\
&= \int_0^\infty \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}\left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1}\exp\left\{\left(-\frac{\delta_0}{2\sigma^2}\right)\right\}\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}m\right\}\frac{|\mathbf{B}_n|^{1/2}}{|\mathbf{B}_0|^{1/2}}d\sigma^2 \\
&= \frac{1}{(2\pi)^{N/2}}\frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}\frac{|\mathbf{B}_n|^{1/2}}{|\mathbf{B}_0|^{1/2}}\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0+N}{2}+1}\exp\left\{\left(-\frac{\delta_0 + m}{2\sigma^2}\right)\right\}d\sigma^2 \\
&= \frac{1}{\pi^{N/2}}\frac{\delta_0^{\alpha_0/2}}{\delta_n^{\alpha_n/2}}\frac{|\mathbf{B}_n|^{1/2}}{|\mathbf{B}_0|^{1/2}}\frac{\Gamma(\alpha_n/2)}{\Gamma(\alpha_0/2)}.
\end{aligned}$$

We can show that $\delta_n = \delta_0 + \mathbf{y}^\top\mathbf{y} + \boldsymbol{\beta}_0^\top\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_n^\top\mathbf{B}_n^{-1}\boldsymbol{\beta}_n = \delta_0 + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top((\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{B}_0)^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ (see Exercise 7). Therefore, if we want to compare two models under this setting, the Bayes factor is

$$\begin{aligned} BF_{12} &= \frac{p(\mathbf{y}|\mathcal{M}_\mathbf{1})}{p(\mathbf{y}|\mathcal{M}_\mathbf{2})} \\ &= \frac{\frac{\delta_{10}^{\alpha_{10}/2}}{\delta_{1n}^{\alpha_{1n}/2}}\frac{|\mathbf{B}_{1n}|^{1/2}}{|\mathbf{B}_{10}|^{1/2}}\frac{\Gamma(\alpha_{1n}/2)}{\Gamma(\alpha_{10}/2)}}{\frac{\delta_{20}^{\alpha_{20}/2}}{\delta_{2n}^{\alpha_{2n}/2}}\frac{|\mathbf{B}_{2n}|^{1/2}}{|\mathbf{B}_{20}|^{1/2}}\frac{\Gamma(\alpha_{2n}/2)}{\Gamma(\alpha_{20}/2)}}, \end{aligned}$$

where subscripts 1 and 2 refer to each model, respectively.

Observe that *ceteris paribus*, the model having better fit, coherence between sample and prior information regarding location parameters, higher prior to posterior precision and less parameters is favored by the Bayes factor. Observe that the Bayes factor rewards model fit as the sum of squared errors is in δ_n, the better fit (lower sum of squared errors), the better the Bayes factor. In addition, a weighted distance between sample and prior location parameters also appears in δ_n, the greater this distance, the worse is model support. The ratio of determinants between posterior and prior covariance matrices is also present, the higher this ratio, the better for the Bayes factor supporting a model due to information gains. To see the effect of model's parsimony, let's take the common situation in applications where $\mathbf{B}_{j0} = c\mathbf{I}_{K_j}$ then $|\mathbf{B}_{j0}| = c^{K_j}$ such that $\left(\frac{|\mathbf{B}_{20}|}{|\mathbf{B}_{10}|}\right)^{1/2} = \left(\frac{c^{K_2/2}}{c^{K_1/2}}\right)$, if $K_2/K_1 > 1$ and $c \to \infty$, the latter implying a non-informative prior, then $BF_{12} \to \infty$, this means infinite evidence supporting the parsimonious model no matter what sample information says. Comparing models having the same number of regressors ($K_1 = K_2$) is not a safe ground as $|\mathbf{B}_0|$ depending on measure units of the regressors such that conclusions regarding model selection depending on this, which is not a nice property. This prevents against using non-informative priors when performing model selection in the Bayesian framework. Observe that this is not the case when $\alpha_0 \to 0$ and $\delta_0 \to 0$, which implies a non-informative prior for the variance parameter.[5] We observe here that $\Gamma(\alpha_{j0})$ cancels out, $\alpha_{jn} \to N$ and $\delta_{jn} \to (\mathbf{y} - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j)^\top(\mathbf{y} - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j) + (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{j0})^\top((\mathbf{X}_j^\top\mathbf{X}_j)^{-1} + \mathbf{B}_{j0})^{-1}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{j0})$, therefore there is not effect. This is due to σ^2 being a common parameter in both models. In general, we can use non-informative priors for common parameters to all models, but we cannot use non-informative priors for non-common parameters when performing model selection using the Bayes factor.

The posterior predictive is equal to

[5] [33] prevents against this common practice.

$$\begin{aligned}\pi(\mathbf{Y}_0|\mathbf{y}) &= \int_0^\infty \int_{R^K} p(\mathbf{Y}_0|\boldsymbol{\beta},\sigma^2,\mathbf{y})\pi(\boldsymbol{\beta}|\sigma^2,\mathbf{y})\pi(\sigma^2|\mathbf{y})d\boldsymbol{\beta}d\sigma^2 \\ &= \int_0^\infty \int_{R^K} p(\mathbf{Y}_0|\boldsymbol{\beta},\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2,\mathbf{y})\pi(\sigma^2|\mathbf{y})d\boldsymbol{\beta}d\sigma^2,\end{aligned}$$

where we take into account independence between $\mathbf{Y}_0$ and $\mathbf{Y}$. Given $\mathbf{X}_0$, which is the $N_0 \times K$ matrix of regressors associated with $\mathbf{Y}_0$, Then,

$$\begin{aligned}\pi(\mathbf{Y}_0|\mathbf{y}) = \int_0^\infty \int_{R^K} \Bigg\{ &(2\pi\sigma^2)^{-\frac{N_0}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta})^\top(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta})^\top\right\} \\ &\times (2\pi\sigma^2)^{-\frac{K}{2}}|\mathbf{B}_n|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\top\mathbf{B}_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\right\} \\ &\times \frac{(\delta_n/2)^{\alpha_n/2}}{\Gamma(\alpha_n/2)}\left(\frac{1}{\sigma^2}\right)^{\alpha_n/2+1}\exp\left\{-\frac{\delta_n}{2\sigma^2}\right\}\Bigg\} d\boldsymbol{\beta}d\sigma^2.\end{aligned}$$

Setting $\mathbf{M} = (\mathbf{X}_0^\top\mathbf{X}_0 + \mathbf{B}_n^{-1})$ and $\boldsymbol{\beta}_* = \mathbf{M}^{-1}(\mathbf{B}_n^{-1}\boldsymbol{\beta}_n + \mathbf{X}_0^\top\mathbf{Y}_0)$, we have $(\mathbf{Y}_0-\mathbf{X}_0\boldsymbol{\beta})^\top(\mathbf{Y}_0-\mathbf{X}_0\boldsymbol{\beta})^\top + (\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\top\mathbf{B}_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n) = (\boldsymbol{\beta}-\boldsymbol{\beta}_*)^\top\mathbf{M}(\boldsymbol{\beta}-\boldsymbol{\beta}_*) + \boldsymbol{\beta}_n^\top\mathbf{B}_n^{-1}\boldsymbol{\beta}_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \boldsymbol{\beta}_*^\top\mathbf{M}\boldsymbol{\beta}_*$. Thus,

$$\begin{aligned}\pi(\mathbf{Y}_0|\mathbf{y}) \propto \int_0^\infty \Bigg\{ &\left(\frac{1}{\sigma^2}\right)^{-\frac{K+N_0+\alpha_n}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}_n^\top\mathbf{B}_n^{-1}\boldsymbol{\beta}_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \boldsymbol{\beta}_*^\top\mathbf{M}\boldsymbol{\beta}_* + \delta_n)\right\} \\ &\times \int_{R^K}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_*)^\top\mathbf{M}(\boldsymbol{\beta}-\boldsymbol{\beta}_*)\right\}d\boldsymbol{\beta}\Bigg\} d\sigma^2,\end{aligned}$$

where the term in the second integral is the kernel of a multivariate normal density with mean $\boldsymbol{\beta}_*$ and covariance matrix $\sigma^2\mathbf{M}^{-1}$. Then,

$$\pi(\mathbf{Y}_0|\mathbf{y}) \propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{N_0+\alpha_n}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}_n^\top\mathbf{B}_n^{-1}\boldsymbol{\beta}_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \boldsymbol{\beta}_*^\top\mathbf{M}\boldsymbol{\beta}_* + \delta_n)\right\} d\sigma^2,$$

which is the kernel of an inverse gamma density. Thus,

$$\pi(\mathbf{Y}_0|\mathbf{y}) \propto \left[\frac{\boldsymbol{\beta}_n^\top\mathbf{B}_n^{-1}\boldsymbol{\beta}_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \boldsymbol{\beta}_*^\top\mathbf{M}\boldsymbol{\beta}_* + \delta_n}{2}\right]^{-\frac{\alpha_n+N_0}{2}}.$$

Setting $\mathbf{C}^{-1} = \mathbf{I}_{N_0} + \mathbf{X}_0\mathbf{B}_n\mathbf{X}_0^\top$ such that $\mathbf{C} = \mathbf{I}_{N_0} - \mathbf{X}_0(\mathbf{B}_n^{-1} +$

$\mathbf{X}_0^\top\mathbf{X}_0)^{-1}\mathbf{X}_0^\top = \mathbf{I}_{N_0} - \mathbf{X}_0\mathbf{M}^{-1}\mathbf{X}_0^\top$,[6] and $\boldsymbol{\beta}_{**} = \mathbf{C}^{-1}\mathbf{X}_0\mathbf{M}^{-1}\mathbf{B}_n^{-1}\boldsymbol{\beta}_n$, then

$$\begin{aligned}\boldsymbol{\beta}_n^\top\mathbf{B}_n^{-1}\boldsymbol{\beta}_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \boldsymbol{\beta}_*^\top\mathbf{M}\boldsymbol{\beta}_* &= \boldsymbol{\beta}_n^\top\mathbf{B}_n^{-1}\boldsymbol{\beta}_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - (\boldsymbol{\beta}_n^\top\mathbf{B}_n^{-1} + \mathbf{Y}_0^\top\mathbf{X}_0)\mathbf{M}^{-1}(\mathbf{B}_n^{-1}\boldsymbol{\beta}_n + \mathbf{X}_0^\top\mathbf{Y}_0)\\ &= \boldsymbol{\beta}_n^\top(\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1}\mathbf{M}^{-1}\mathbf{B}_n^{-1})\boldsymbol{\beta}_n + \mathbf{Y}_0^\top\mathbf{C}\mathbf{Y}_0\\ &\quad - 2\mathbf{Y}_0^\top\mathbf{C}\mathbf{C}^{-1}\mathbf{X}_0\mathbf{M}^{-1}\mathbf{B}_n^{-1}\boldsymbol{\beta}_n + \boldsymbol{\beta}_{**}^\top\mathbf{C}\boldsymbol{\beta}_{**} - \boldsymbol{\beta}_{**}^\top\mathbf{C}\boldsymbol{\beta}_{**}\\ &= \boldsymbol{\beta}_n^\top(\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1}\mathbf{M}^{-1}\mathbf{B}_n^{-1})\boldsymbol{\beta}_n + (\mathbf{Y}_0 - \boldsymbol{\beta}_{**})^\top\mathbf{C}(\mathbf{Y}_0 - \boldsymbol{\beta}_{**})\\ &\quad - \boldsymbol{\beta}_{**}^\top\mathbf{C}\boldsymbol{\beta}_{**},\end{aligned}$$

where $\boldsymbol{\beta}_n^\top(\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1}\mathbf{M}^{-1}\mathbf{B}_n^{-1})\boldsymbol{\beta}_n = \boldsymbol{\beta}_{**}^\top\mathbf{C}\boldsymbol{\beta}_{**}$ and $\boldsymbol{\beta}_{**} = \mathbf{X}_0\boldsymbol{\beta}_n$ (see Exercise 8).

Then,

$$\begin{aligned}\pi(\mathbf{Y}_0|\mathbf{y}) &\propto \left[\frac{(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}_n)^\top\mathbf{C}(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}_n) + \delta_n}{2}\right]^{-\frac{\alpha_n+N_0}{2}}\\ &\propto \left[\frac{(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}_n)^\top\left(\frac{\mathbf{C}\alpha_n}{\delta_n}\right)(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}_n)}{\alpha_n} + 1\right]^{-\frac{\alpha_n+N_0}{2}}.\end{aligned}$$

The posterior predictive is a multivariate t distribution, $\mathbf{Y}_0|\mathbf{y} \sim t\left(\mathbf{X}_0\boldsymbol{\beta}_n, \frac{\delta_n(\mathbf{I}_{N_0}+\mathbf{X}_0\mathbf{B}_n\mathbf{X}_0^\top)}{\alpha_n}, \alpha_n\right)$ centered at $\mathbf{X}_0\boldsymbol{\beta}_n$.

Example: Demand of electricity

We study in this example the determinants of monthly demand of electricity by Colombian households. There is information of 2103 households, particularly, average price (USD/kWh), indicators of socioeconomic conditions of the neighborhood where the household is located (IndSocio1 is the lowest and IndSocio3 is the highest), an indicator if the household is located in a municipality that is above 1000 meters above the sea level, the number of rooms in the house, the number of members of the households, presence of children in the household (1 is yes), and monthly income (USD). The specification is

$$\begin{aligned}\log(\text{Electricity}_i) = {}& \beta_1\log(\text{price}_i) + \beta_2\text{IndSocio1}_i + \beta_3\text{IndSocio2}_i + \beta_4\text{Altitude}_i\\ &+ \beta_5\text{Nrooms}_i + \beta_6\text{HouseholdMem}_i + \beta_7\text{Children}_i\\ &+ \beta_8\log(\text{Income}_i) + \beta_9 + \mu.\end{aligned}$$

We use a non-informative vague prior setting such that $\alpha_0 = \delta_0 = 0.001$, $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\mathbf{B}_0 = c_0\mathbf{I}_k$, where $c_0 = 1000$ and k is the number of regressors.

The results from the R code (see below) is that the posterior mean of the own-price of electricity demand is -1.09, and the 95% symmetric credible interval is (-1.47, -0.71). Households in neighborhoods of low socioeconomic conditions and located in municipalities 1000 meters above the sea level consume less electricity, 32.7% and 19.7% on average, respectively. An additional

[6] Using $(\mathbf{A}+\mathbf{BDC})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}^{-1} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$

room implies a 8.7% increase in electricity consumption, and another household member increases consumption in 5.9% on average. The income elasticity mean estimate is 0.074, which means that 10% increase of income increases 0.74% electricity demand.

We want to check the results of the Bayes factor comparing the previous specification (model 1) with other specification without considering the price of electricity (model 2), that is,

$$\begin{aligned}\log(\text{Electricity}_i) = {} & \beta_1 \text{IndSocio1}_i + \beta_2 \text{IndSocio2}_i + \beta_3 \text{Altitude}_i + \beta_4 \text{Nrooms}_i \\ & + \beta_5 \text{HouseholdMem}_i + \beta_6 \text{Children}_i + \beta_7 \log(\text{Income}_i) \\ & + \beta_8 + \mu.\end{aligned}$$

In particular, we check what happens as c_0 increases from 10^0 to 10^{20}. We see that when $c_0 = 1$, $BF_{12} = 8.68 \times 10^{+16}$, which means very strong evidence in favor of the model including the price of electricity. However, as c_0 increases, the Bayes factor decreases, which means evidence supporting the model 2, for instance, $BF_{12} = 3.11 \times 10^{-4}$ when $c_0 = 10^{20}$. This is an example of the problem of using non-informative priors to calculate the Bayes factor; there is very strong evidence to support the parsimonious model when $c_0 \to \infty$.

We can get the posterior predictive distribution of the monthly electricity demand of a household located in the lowest socioeconomic condition in a municipality located below 1000 meters above the sea level, 2 rooms, 3 members with children, a monthly income equal to USD 500, and an electricity price equal to USD/kWh 0.15. Figure 4.1 shows the histogram of the predictive posterior distribution, the highest posterior density credible interval at 95% is between kWh 44.4 and kWh 373.9, and the posterior mean is kWh 169.4.

R code. Demand of electricity, posterior predictive distribution

```
rm(list = ls())
set.seed(010101)
# Electricity demand
DataUt <- read.csv("DataApplications/Utilities.csv", sep = ",", header = TRUE, fileEncoding = "latin1")
DataUtEst <- DataUt %>%
  filter(Electricity != 0)
attach(DataUtEst)
# Dependent variable: Monthly consumption (kWh) in log
Y <- log(Electricity)
# Regressors quantity including intercept
X <- cbind(LnPriceElect, IndSocio1, IndSocio2, Altitude, Nrooms, HouseholdMem, Children, Lnincome, 1)
# LnPriceElect: Price per kWh (USD) in log
# IndSocio1, IndSocio2, IndSocio3: Indicators socio-economic condition (1) is the lowest and (3) the highest
# Altitude: Indicator of household location (1 is more than 1000 meters above sea level)
# Nrooms: Number of rooms in house
# HouseholdMem: Number of household members
# Children: Indicator por presence of children in household (1)
# Lnincome: Monthly income (USD) in log
k <- dim(X)[2]
N <- dim(X)[1]
# Hyperparameters
d0 <- 0.001/2
a0 <- 0.001/2
b0 <- rep(0, k)
B0 <- 1000*diag(k)
# Posterior parameters
bhat <- solve(t(X)%*%X)%*%t(X)%*%Y
Bn <- as.matrix(Matrix::forceSymmetric(solve(solve(B0) + t(X)%*%X))) # Force this matrix to be symmetric
bn <- Bn%*%(solve(B0)%*%b0 + t(X)%*%X%*%bhat)
dn <- as.numeric(d0 + t(Y)%*%Y+t(b0)%*%solve(B0)%*%b0-t(bn)%*%solve(Bn)%*%bn)
an <- a0 + N
Hn <- Bn*dn/an
# Posterior draws
S <- 10000 # Number of draws from posterior distributions
sig2 <- MCMCpack::rinvgamma(S,an/2,dn/2)
summary(coda::mcmc(sig2))
```

R code. Demand of electricity, posterior distribution

```
Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each
variable, plus standard error of the mean:

Mean             SD       Naive SE  Time-series SE
2.361e-01      7.617e-03  7.617e-05 7.617e-05

2. Quantiles for each variable:

2.5%    25%    50%    75%  97.5%
0.2217 0.2309 0.2360 0.2412 0.2513

Betas <- LaplacesDemon::rmvt(S, bn, Hn, an)
summary(coda::mcmc(Betas))
Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each
variable, plus standard error of the mean:

              Mean       SD    Naive SE  Time-series SE
LnPriceElect -1.09043 0.19459 0.0019459 0.0019459
IndSocio1    -0.32783 0.05294 0.0005294 0.0005294
IndSocio2    -0.05737 0.04557 0.0004557 0.0004557
Altitude     -0.19780 0.02386 0.0002386 0.0002429
Nrooms        0.08731 0.01094 0.0001094 0.0001119
HouseholdMem  0.05987 0.01334 0.0001334 0.0001334
Children      0.05696 0.03043 0.0003043 0.0003043
Lnincome      0.07447 0.01223 0.0001223 0.0001223
              2.52296 0.35077 0.0035077 0.0035077

2. Quantiles for each variable:

                  2.5%      25%      50%      75%    97.5%
LnPriceElect -1.472069 -1.22432 -1.08961 -0.95703 -0.71429
IndSocio1    -0.435957 -0.36228 -0.32731 -0.29133 -0.22588
IndSocio2    -0.147252 -0.08744 -0.05757 -0.02650  0.03254
Altitude     -0.244759 -0.21372 -0.19783 -0.18164 -0.15094
Nrooms        0.066432  0.07985  0.08709  0.09480  0.10864
HouseholdMem  0.033623  0.05089  0.05975  0.06889  0.08596
Children     -0.002259  0.03637  0.05698  0.07736  0.11681
Lnincome      0.050536  0.06614  0.07449  0.08283  0.09852
              1.835507  2.28703  2.52165  2.76364  3.21199

```

R code. Demand of electricity, Bayes factor

```
# Log marginal function (multiply by -1 due to minimization)
LogMarLikLM <- function(X, c0){
  k <- dim(X)[2]
  N <- dim(X)[1]
  # Hyperparameters
  B0 <- c0*diag(k)
  b0 <- rep(0, k)
  # Posterior parameters
  bhat <- solve(t(X)%*%X)%*%t(X)%*%Y
  # Force this matrix to be symmetric
  Bn <- as.matrix(Matrix::forceSymmetric(solve(solve(B0) + t
    (X)%*%X)))
  bn <- Bn%*%(solve(B0)%*%b0 + t(X)%*%X%*%bhat)
  dn <- as.numeric(d0 + t(Y)%*%Y+t(b0)%*%solve(B0)%*%b0-t(bn
    )%*%solve(Bn)%*%bn)
  an <- a0 + N
  # Log marginal likelihood
  logpy <- (N/2)*log(1/pi)+(a0/2)*log(d0)-(an/2)*log(dn) +
    0.5*log(det(Bn)/det(B0)) + lgamma(an/2)-lgamma(a0/2)
  return(-logpy)
}
cs <- c(10^0, 10^3, 10^6, 10^10, 10^12, 10^15, 10^20)
# Observe -1 to recover the right sign
LogML <- sapply(cs, function(c) {-LogMarLikLM(c0=c, X = X)})
# Regressor without price
Xnew <- cbind(IndSocio1, IndSocio2, Altitude, Nrooms,
    HouseholdMem, Children, Lnincome, 1)
# Observe -1 to recover the right sign
LogMLnew <- sapply(cs, function(c) {-LogMarLikLM(c0=c,X =
    Xnew)})
# Bayes factor
BF <- exp(LogML - LogMLnew)
BF
8.687567e+16 1.006679e+05 3.108415e+03 3.108340e+01 3.108343
    e+00 9.829443e-02 3.108343e-04
# Empirical Bayes: Obtain c0 maximizing the log
marginal likelihood
c0 <- c0
EB <- optim(c0, fn = LogMarLikLM, method = "Brent", lower =
    0.0001, upper = 10^6, X = X)
EB$par
3.254822
EB$value
1404.108
EBnew <- optim(c0, fn = LogMarLikLM, method = "Brent", lower
     = 0.0001, upper = 10^6, X = Xnew)
EBnew$par
10.00597
EBnew$value
1422.199
# Change of order to take into account the -1 in the
    LogMarLikLM function
BFEM <- exp(EBnew$value - EB$value)
BFEM
71897938
```

R code. Demand of electricity, predictive distribution

```
# Predictive distribution
Xpred <- c(log(0.15), 1, 0, 0, 2, 3, 1, log(500), 1)
Mean <- Xpred%*%bn
Hn <- dn*(1+t(Xpred)%*%Bn%*%Xpred)/an
ExpKwH <- exp(LaplacesDemon::rmvt(S, Mean, Hn, an))
summary(ExpKwH)
Min.    :  24.06
1st Qu.: 121.70
Median : 169.37
Mean   : 189.60
3rd Qu.: 234.19
Max.   :1243.68
HDI <- HDInterval::hdi(ExpKwH, credMass = 0.95) # Highest
    posterior density credible interval
HDI
lower  44.40203
upper 373.86494
hist(ExpKwH, main = "Histogram: Monthly demand of
    electricity", xlab = "Monthly kWh", col = "blue", breaks
     = 50)
```

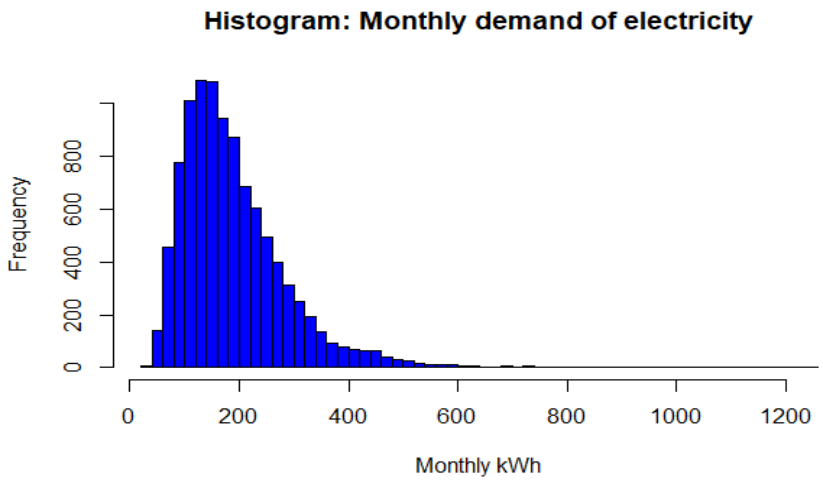

FIGURE 4.1
Histogram using the posterior predictive distribution of electricity demand

4.4 Multivariate linear regression: The conjugate normal-normal/inverse Wishart model

Let's study the multivariate regression setting where there are N-dimensional vectors $\mathbf{y}_m$, $m = 1, 2, \ldots, M$ such that $\mathbf{y}_m = \mathbf{X}\boldsymbol{\beta}_m + \mu_m$, $\mathbf{X}$ is the set of common regressors, and μ_m is the N-dimensional vector of stochastic errors for each equation such that $\mathbf{U} = [\mu_1\ \mu_2\ \ldots\ \mu_M] \sim MN_{N,M}(\mathbf{0}, \mathbf{I}_N, \boldsymbol{\Sigma})$, that is, a matrix variate normal distribution where $\boldsymbol{\Sigma}$ is the covariance matrix of each i-th row of $\mathbf{U}$, $i = 1, 2, \ldots, N$, and we are assuming independence between the rows. Then, $vec(\mathbf{U}) \sim N_{N\times M}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I_N})$.[7]

This framework can be written in matrix form

$$\underbrace{\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2M} \\ \vdots & \vdots & \ldots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NM} \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ldots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\ \vdots & \vdots & \ldots & \vdots \\ \beta_{K1} & \beta_{K2} & \cdots & \beta_{KM} \end{bmatrix}}_{\mathbf{B}}$$
$$+ \underbrace{\begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1M} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2M} \\ \vdots & \vdots & \ldots & \vdots \\ \mu_{N1} & \mu_{N2} & \cdots & \mu_{NM} \end{bmatrix}}_{\mathbf{U}}.$$

Therefore, $\mathbf{Y} \sim N_{N\times M}(\mathbf{XB}, \boldsymbol{\Sigma} \otimes \mathbf{I_N})$,[8]

$$\begin{aligned} p(\mathbf{Y}|\mathbf{B}, \boldsymbol{\Sigma}, \mathbf{X}) &\propto |\boldsymbol{\Sigma}|^{-N/2} \exp\left\{-\frac{1}{2} tr\left[(\mathbf{Y} - \mathbf{XB})^\top (\mathbf{Y} - \mathbf{XB})\boldsymbol{\Sigma}^{-1}\right]\right\} \\ &= |\boldsymbol{\Sigma}|^{-N/2} \exp\left\{-\frac{1}{2} tr\left[\left(\mathbf{S} + (\mathbf{B} - \widehat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X}(\mathbf{B} - \widehat{\mathbf{B}})\right)\boldsymbol{\Sigma}^{-1}\right]\right\}, \end{aligned}$$

where $\mathbf{S} = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^\top (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})$, $\widehat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$ (see Exercise 9).

The conjugate prior for this models is $\pi(\mathbf{B}, \boldsymbol{\Sigma}) = \pi(\mathbf{B}|\boldsymbol{\Sigma})\pi(\boldsymbol{\Sigma})$ where $\pi(\mathbf{B}|\boldsymbol{\Sigma}) \sim N_{K\times M}(\mathbf{B}_0, \mathbf{V}_0, \boldsymbol{\Sigma})$ and $\pi(\boldsymbol{\Sigma}) \sim IW(\boldsymbol{\Psi}_0, \alpha_0)$, that is,

[7] *vec* denotes the vectorization operation, and $\otimes$ denotes the kronecker product.

[8] We can write down the former expression in a more familiar way using vectorization properties, $\underbrace{vec(Y)}_{\mathbf{y}} = \underbrace{(\mathbf{I}_M \otimes \mathbf{X})}_{\mathbf{Z}} \underbrace{vec(\mathbf{B})}_{\boldsymbol{\beta}} + \underbrace{vec(\mathbf{U})}_{\mu}$, where $\mathbf{y} \sim N_{N\times M}(\mathbf{Z}\boldsymbol{\beta}, \boldsymbol{\Sigma} \otimes \mathbf{I_N})$.

$$\pi(\mathbf{B},\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-K/2} \exp\left\{-\frac{1}{2} tr\left[(\mathbf{B}-\mathbf{B}_0)^\top \mathbf{V}_0^{-1}(\mathbf{B}-\mathbf{B}_0)\boldsymbol{\Sigma}^{-1}\right]\right\}$$
$$\times |\boldsymbol{\Sigma}|^{-(\alpha_0+M+1)/2} \exp\left\{-\frac{1}{2} tr\left[\boldsymbol{\Psi}_0\boldsymbol{\Sigma}^{-1}\right]\right\}.$$

The posterior distribution is given by

$$\begin{aligned}\pi(\mathbf{B},\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}) &\propto p(\mathbf{Y}|\mathbf{B},\boldsymbol{\Sigma},\mathbf{X})\pi(\mathbf{B}|\boldsymbol{\Sigma})\pi(\boldsymbol{\Sigma}) \\ &\propto |\boldsymbol{\Sigma}|^{-\frac{N+K+\alpha_0+M+1}{2}} \\ &\times \exp\left\{-\frac{1}{2} tr\left[(\boldsymbol{\Psi_0}+\mathbf{S}+(\mathbf{B}-\mathbf{B_0})^\top \mathbf{V_0^{-1}}(\mathbf{B}-\mathbf{B_0})\right.\right. \\ &\left.\left.+(\mathbf{B}-\widehat{\mathbf{B}})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{B}-\widehat{\mathbf{B}}))\boldsymbol{\Sigma}^{-1}\right]\right\}.\end{aligned}$$

Completing the squares on $\mathbf{B}$ and collecting the remaining terms in the bracket yields

$$\boldsymbol{\Psi}_0+\mathbf{S}+(\mathbf{B}-\mathbf{B}_0)^\top\mathbf{V}_0^{-1}(\mathbf{B}-\mathbf{B}_0)+(\mathbf{B}-\widehat{\mathbf{B}})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{B}-\widehat{\mathbf{B}}) = (\mathbf{B}-\mathbf{B}_n)^\top\mathbf{V}_n^{-1}(\mathbf{B}-\mathbf{B}_n)+\boldsymbol{\Psi}_n,$$

where

$$\begin{aligned}\mathbf{B}_n &= (\mathbf{V}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{V}_0^{-1}\mathbf{B}_0+\mathbf{X}^\top\mathbf{Y}) = (\mathbf{V}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{V}_0^{-1}\mathbf{B}_0+\mathbf{X}^\top\mathbf{X}\widehat{\mathbf{B}}), \\ \mathbf{V}_n &= (\mathbf{V}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}, \\ \boldsymbol{\Psi}_n &= \boldsymbol{\Psi}_0+\mathbf{S}+\mathbf{B}_0^\top\mathbf{V}_0^{-1}\mathbf{B}_0+\widehat{\mathbf{B}}^\top\mathbf{X}^\top\mathbf{X}\widehat{\mathbf{B}}-\mathbf{B}_n^\top\mathbf{V}_n^{-1}\mathbf{B}_n.\end{aligned}$$

Thus, the posterior distribution can be written as

$$\pi(\mathbf{B},\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}) \propto |\boldsymbol{\Sigma}|^{-K/2} \exp\left\{-\frac{1}{2} tr\left[(\mathbf{B}-\mathbf{B}_n)^\top \mathbf{V}_n^{-1}(\mathbf{B}-\mathbf{B}_n)\boldsymbol{\Sigma}^{-1}\right]\right\}$$
$$\times |\boldsymbol{\Sigma}|^{-\frac{N+\alpha_0+M+1}{2}} \exp\left\{-\frac{1}{2} tr\left[\boldsymbol{\Psi}_n\boldsymbol{\Sigma}^{-1}\right]\right\}.$$

That is $\pi(\mathbf{B},\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}) = \pi(\mathbf{B}|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X})\pi(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X})$ where $\pi(\mathbf{B}|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X}) \sim N_{K\times M}(\mathbf{B}_n,\mathbf{V}_n,\boldsymbol{\Sigma})$ and $\pi(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}) \sim IW(\boldsymbol{\Psi}_n,\alpha_n)$, $\alpha_n = N+\alpha_0$. Observe again that we can write down the posterior mean as a weighted average between prior and sample information such that $\mathbf{V}_0 \to \infty$ implies $\mathbf{B}_n \to \hat{\mathbf{B}}$, as we show in the univariate linear model.

The marginal posterior for $\mathbf{B}$ is given by

$$\begin{aligned}\pi(\mathbf{B}|\mathbf{Y},\mathbf{X}) &\propto \int_{\mathcal{S}} |\boldsymbol{\Sigma}|^{-(\alpha_n+K+M+1)/2} \\ &\times \exp\left\{-\frac{1}{2} tr\left\{\left[(\mathbf{B}-\mathbf{B}_n)^\top\mathbf{V}_n^{-1}(\mathbf{B}-\mathbf{B}_n)+\boldsymbol{\Psi}_n\right]\boldsymbol{\Sigma}^{-1}\right\}\right\} d\boldsymbol{\Sigma} \\ &\propto |(\mathbf{B}-\mathbf{B}_n)^\top\mathbf{V}_n^{-1}(\mathbf{B}-\mathbf{B}_n)+\boldsymbol{\Psi}_n|^{-(K+\alpha_n)/2} \\ &= \left[|\boldsymbol{\Psi}_n| \times |\mathbf{I}_K+\mathbf{V}_n^{-1}(\mathbf{B}-\mathbf{B}_n)\boldsymbol{\Psi}_n^{-1}(\mathbf{B}-\mathbf{B}_n)^\top|\right]^{-(\alpha_n+1-M+K+M-1)/2} \\ &\propto |\mathbf{I}_K+\mathbf{V}_n^{-1}(\mathbf{B}-\mathbf{B}_n)\boldsymbol{\Psi}_n^{-1}(\mathbf{B}-\mathbf{B}_n)^\top|^{-(\alpha_n+1-M+K+M-1)/2}.\end{aligned}$$

The second line uses the inverse Wishart distribution, the third line the Sylverter's theorem, and the last line is the kernel of a matrix t distribution, that is, $\mathbf{B}|\mathbf{Y}, \mathbf{X} \sim T_{K\times M}(\mathbf{B}_n, \mathbf{V}_n, \mathbf{\Psi}_n)$ with $\alpha_n + 1 - M$ degrees of freedom.

Observe that $vec(\mathbf{B})$ has mean $vec(\mathbf{B}_n)$ and variance $(\mathbf{V}_n \otimes \mathbf{\Psi}_n)/(\alpha_n - M - 1)$ based on its marginal distribution. On the other hand, the variance based on the conditional distribution is $\mathbf{V}_n \otimes \mathbf{\Sigma}$, where the mean of $\mathbf{\Sigma}$ is $\mathbf{\Psi}_n/(\alpha_n - M - 1)$.

The marginal likelihood is the following,

$$
\begin{aligned}
p(\mathbf{Y}) &= \int_{\mathcal{B}}\int_{\mathcal{S}} \left\{ (2\pi)^{-NM/2} |\mathbf{\Sigma}|^{-N/2} \exp\left\{ -\frac{1}{2} tr\left[\mathbf{S} + (\mathbf{B}-\widehat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{B}-\widehat{\mathbf{B}}) \right] \mathbf{\Sigma}^{-1} \right\} \right. \\
&\quad \times (2\pi)^{-KM/2} |\mathbf{V}_0|^{-M/2} |\mathbf{\Sigma}|^{-K/2} \exp\left\{ -\frac{1}{2} tr\left[(\mathbf{B}-\mathbf{B}_0)^\top \mathbf{V}_0^{-1} (\mathbf{B}-\mathbf{B}_0) \mathbf{\Sigma}^{-1} \right] \right\} \\
&\quad \left. \times \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} |\mathbf{\Sigma}|^{-(\alpha_0+M+1)/2} \exp\left\{ -\frac{1}{2} tr\left[\mathbf{\Psi}_0 \mathbf{\Sigma}^{-1} \right] \right\} \right\} d\mathbf{\Sigma} d\mathbf{B} \\
&= (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} \\
&\quad \times \int_{\mathcal{B}}\int_{\mathcal{S}} \left\{ |\mathbf{\Sigma}|^{-(\alpha_0+N+K+M+1)/2} \right. \\
&\quad \left. \exp\left\{ -\frac{1}{2} tr\left[\mathbf{S} + (\mathbf{B}-\widehat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{B}-\widehat{\mathbf{B}}) + (\mathbf{B}-\mathbf{B}_0)^\top \mathbf{V}_0^{-1} (\mathbf{B}-\mathbf{B}_0) + \mathbf{\Psi}_0 \right] \mathbf{\Sigma}^{-1} \right\} \right\} d\mathbf{\Sigma} d\mathbf{B} \\
&= (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2) \\
&\quad \times \int_{\mathcal{B}} \left| \mathbf{S} + (\mathbf{B}-\widehat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{B}-\widehat{\mathbf{B}}) + (\mathbf{B}-\mathbf{B}_0)^\top \mathbf{V}_0^{-1} (\mathbf{B}-\mathbf{B}_0) + \mathbf{\Psi}_0 \right|^{-(\alpha_n+K)/2} d\mathbf{B} \\
&= (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2) \\
&\quad \times \int_{\mathcal{B}} \left| (\mathbf{B}-\widehat{\mathbf{B}}_n)^\top \mathbf{V}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n) + \mathbf{\Psi}_n \right|^{-(\alpha_n+K)/2} d\mathbf{B} \\
&= (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2) \\
&\quad \times \int_{\mathcal{B}} \left[|\mathbf{\Psi}_n| \times |\mathbf{I}_K + \mathbf{V}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n) \mathbf{\Psi}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n)^\top | \right]^{-(\alpha_n+K)/2} d\mathbf{B} \\
&= |\mathbf{\Psi}_n|^{-(\alpha_n+K)/2} (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2)}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} \\
&\quad \times \int_{\mathbf{B}} \left| \mathbf{I}_K + \mathbf{V}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n) \mathbf{\Psi}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n)^\top \right|^{-(\alpha_n+1-M+K+M-1)/2} d\mathbf{B} \\
&= |\mathbf{\Psi}_n|^{-(\alpha_n+K)/2} (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2)}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} \\
&\quad \times \pi^{MK/2} \frac{\Gamma_M((\alpha_n+1-M+M-1)/2)}{\Gamma_M((\alpha_n+1-M+K+M-1)/2)} |\mathbf{\Psi}_n|^{K/2} |\mathbf{V}_n|^{M/2} \\
&= \frac{|\mathbf{V}_n|^{M/2}}{|\mathbf{V}_0|^{M/2}} \frac{|\mathbf{\Psi}_0|^{\alpha_0/2}}{|\mathbf{\Psi}_n|^{\alpha_n/2}} \frac{\Gamma_M(\alpha_n/2)}{\Gamma_M(\alpha_0/2)} \pi^{-MN/2}.
\end{aligned}
$$

The third equality follows from having the kernel of a inverse Wishart distribution, the fifth from the Silvester's theorem, and the seventh from having the kernel of a matrix t distribution.

Observe that this last expression is the multivariate case of the marginal likelihood of the univariate regression model. Taking into account that

$$\begin{aligned}(\mathbf{A}+\mathbf{B})^{-1} &= \mathbf{A}^{-1} - (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{A}^{-1} \\ &= \mathbf{B}^{-1} - (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1} \\ &= \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{B}^{-1},\end{aligned}$$

we can show that $\boldsymbol{\Psi}_n = \boldsymbol{\Psi}_0 + \mathbf{S} + (\hat{\mathbf{B}} - \mathbf{B}_0)^\top \mathbf{V}_n (\hat{\mathbf{B}} - \mathbf{B}_0)$ (see Exercise 7). Therefore, the marginal likelihood rewards fit (smaller sum of squares, $\mathbf{S}$), similarity between prior and sample information regarding location parameters, and information gains in variability from $\mathbf{V}_0$ to $\mathbf{V}_n$.

Given a matrix of regressors $\mathbf{X}_0$ for N_0 unobserved units, the predictive density of $\mathbf{Y}_0$ given $\mathbf{Y}$, $\pi(\mathbf{Y}_0|\mathbf{Y})$ is a matrix t distribution $T_{N_0,M}(\alpha_n - M + 1, \mathbf{X}_0\mathbf{B}_n, \mathbf{I}_{N_0} + \mathbf{X}_0\mathbf{V}_n\mathbf{X}_0^\top, \boldsymbol{\Psi}_n)$ (see Exercise 6). Observe that the prediction is centered at $\mathbf{X}_0\mathbf{B}_n$, and the covariance matrix of $vec(\mathbf{Y}_0)$ is $\frac{(\mathbf{I}_{N_0}+\mathbf{X}_0\mathbf{V}_n\mathbf{X}_0^\top)\otimes\boldsymbol{\Psi}_n}{\alpha_n - M - 1}$.

4.5 Summary

We introduce the conjugate family models for discrete and continuous data. These models are the basic Bayesian framework to due its mathematical tractability as we get closed-form expressions for the posterior distributions, the marginal likelihood, and the predictive distribution. We also present the Bayesian linear univariate and multivariate regression frameworks under conjugate families. This is the cornerstone to perform regression analysis in the Bayesian setting.

4.6 Exercises

1. Write in the canonical form the distribution of the Bernoulli example, and find the mean and variance of the sufficient statistic.
2. Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from N *binomial experiments* each having known size n_i and same unknown probability θ. Show that $p(\mathbf{y}|\theta)$ is in the exponential family, and find the posterior distribution, the marginal likelihood and the predictive distribution of the binomial-beta model assuming the number of trials is known.
3. Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a *exponential distribution*. Show that $p(\mathbf{y}|\lambda)$ is in the exponential family, and find

the posterior distribution, marginal likelihood and predictive distribution of the exponential-gamma model.

4. Given $\mathbf{y} \sim N_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, that is, a *multivariate normal distribution* show that $p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is in the exponential family.
5. Find the marginal likelihood in the normal/inverse-Wishart model.
6. Find the posterior predictive distribution in the normal/inverse-Wishart model, and show that $\mathbf{Y}_0|\mathbf{Y} \sim T_{N_0,M}(\alpha_n - M + 1, \mathbf{X}_0\mathbf{B}_n, \mathbf{I}_{N_0} + \mathbf{X}_0\mathbf{V}_n\mathbf{X}_0^\top, \boldsymbol{\Psi}_n)$.
7. Show that $\delta_n = \delta_0 + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top((\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{B}_0)^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ in the linear regression model, and that $\boldsymbol{\Psi}_n = \boldsymbol{\Psi}_0 + \mathbf{S} + (\hat{\mathbf{B}} - \mathbf{B}_0)^\top\mathbf{V}_n(\hat{\mathbf{B}} - \mathbf{B}_0)$ in the linear multivariate regression model.
8. Show that in the linear regression model $\boldsymbol{\beta}_n^\top(\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1}\mathbf{M}^{-1}\mathbf{B}_n^{-1})\boldsymbol{\beta}_n = \boldsymbol{\beta}_{**}^\top\mathbf{C}\boldsymbol{\beta}_{**}$ and $\boldsymbol{\beta}_{**} = \mathbf{X}_0\boldsymbol{\beta}_n$.
9. Show that $(\mathbf{Y} - \mathbf{X}\mathbf{B})^\top(\mathbf{Y} - \mathbf{X}\mathbf{B}) = \mathbf{S} + (\mathbf{B} - \widehat{\mathbf{B}})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{B} - \widehat{\mathbf{B}})$ where $\mathbf{S} = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})$, $\widehat{\mathbf{B}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ in the multivariate regression model.
10. **What is the probability that the Sun will rise tomorrow?**

    This is the most famous Richard Price's example developed in the Appendix of the Bayes' theorem paper [6]. Here, we implicitly use *Laplace's Rule of Succession* to solve this question. In particular, if we were a priori uncertain about the probability the Sun will rise on a specified day, we can assume a prior uniform distribution over (0,1), that is, a beta (1,1) distribution. Then, what is the probability that the Sun will rise?
11. Using information from Public Policy Polling in September 27th-28th for the 2016 presidential five-way race in USA, there are 411, 373 and 149 sampled people supporting Hillary Clinton, Donald Trump and other, respectively.

    - Find the posterior probability of the percentage difference of people supporting Hillary versus Trump according to this data using a non-informative prior, that is, $\alpha_0 = [1\ 1\ 1]$ in the multinomial-Dirichlet model. What is the probability of having more supports of Hillary vs Trump?
    - What is the probability that sampling one hundred independent individuals 44, 40 and 16 support Hillary, Trump and other, respectively?

12. **Math test example continues**

    You have a random sample of math scores of size $N = 50$ from a normal distribution, $Y_i \sim \mathcal{N}(\mu, \sigma)$. The sample mean and variance are

equal to 102 and 10, respectively. Using the normal-normal/inverse-gamma model where $\mu_0 = 100$, $\beta_0 = 1$, $\alpha_0 = \delta_0 = 0.001$

- Get a 95% confidence and credible interval for μ.
- What is the posterior probability that $\mu > 103$?

13. **Demand of electricity example continues**

Set c_0 such that maximizes the marginal likelihood in the specifications with and without electricity price in the example of demand of electricity (empirical Bayes). Then, calculate the Bayes factor, and conclude if there is evidence supporting the inclusion of the price of electricity in the demand equation.

14. **Utility demand**

Use the file *Utilities.csv* to estimate a multivariate linear regression model where $\mathbf{Y}_i = [\log(\text{electricity}_i)\ \log(\text{water}_i)\ \log(\text{gas}_i)]$ as function of $\log(\text{electricity price}_i)$, $\log(\text{water price}_i)$, $\log(\text{gas price}_i)$, IndSocio1_i, IndSocio2_i, Altitude_i, Nrooms_i, HouseholdMem_i, Children_i, and $\log(\text{Income}_i)$. Set a non-informative prior framework, $\mathbf{B}_0 = [0]_{11\times 3}$, $\mathbf{V}_0 = 1000\mathbf{I}_{11}$, $\mathbf{\Psi}_0 = 1000\mathbf{I}_3$ and $\alpha_0 = 3$, where we have $K = 11$ (regressors plus intercept) and $M = 3$ (equations) in this exercise.

(a) Find the posterior mean estimates and the highest posterior density intervals at 95% of $\mathbf{B}$ and $\mathbf{\Sigma}$. Use the marginal distribution and the conditional distribution to obtain the posterior estimates of $\mathbf{B}$, and compare the results.

(b) Find the Bayes factor comparing the baseline model in this exercise with the same specification but using the income in dollars. Now, calculate the Bayes factor using the income in thousand dollars. Is there any difference?

(c) Find the predictive distribution for the monthly demand of electricity, water and gas in the baseline specification of a household located in the lowest socioeconomic condition in a municipality located below 1000 meters above the sea level, 2 rooms, 3 members with children, a monthly income equal to USD 500, an electricity price equal to USD/kWh 0.15, a water price equal to USD/M^3 0.70, and a gas price equal to USD/M^3 0.75.

5

Simulation methods

Part II

Regression models: A GUIded tour

6

Graphical user interface

This chapter presents our graphical user interface (GUI) to carry out Bayesian regression analysis in a very friendly environment without any programming skills (drag and drop). Our GUI is based on an interactive web application using *shiny*, and packages from **R** software, and is designed for teaching and applied purposes at an introductory level. In the next chapters of the second part of this book we carry out some applications to highlight the potential of our GUI for applied researchers and practitioners.

6.1 Introduction

Our GUI allows performing inference using Bayesian regression analysis without requiring programming skills. The latter seems to be a significant impediment to increasing the use of the Bayesian framework [91, 47].

There are other available graphical user interfaces for carrying out Bayesian regression analysis. *ShinyStan* [85] is a very flexible open source program, but users are required to have some programming skills. *BugsXLA* [91] is open source, but less flexible. However, users do not need to have programming skills. *Bayesian regression: Nonparametric and parametric models* [47] is a very flexible and friendly GUI that is based on *MATLAB Compiler* for a 64-bit Windows computer. Its focus is on Bayesian nonparametric regressions, and it can be thought of for users who have mastered basic parametric models, such as the ones that we show in our GUI. There are also *MATLAB toolkit*, *Stata* and *BayES*, but these are not open sources.

We developed our GUI based on an interactive web application using shiny [16], and some libraries in **R** [69]. The specific libraries and commands that are used in our GUI can be seen in Table 15.1. It has ten univariate models, four multivariate, time series models, three hierarchical longitudinal, and seven Bayesian model averaging frameworks. In addition, it gives basic summaries and diagnostics of the posterior chains, as well as the posterior chains themselves, and different plots, such as trace, autocorrelation and densities.

In terms of its flexibility and possibilities, our GUI lies between *ShinyStan* and *BugsXLA*: users are not required to have any programming skills, but it is not as advanced as [47]'s software. However, our GUI can be run in any

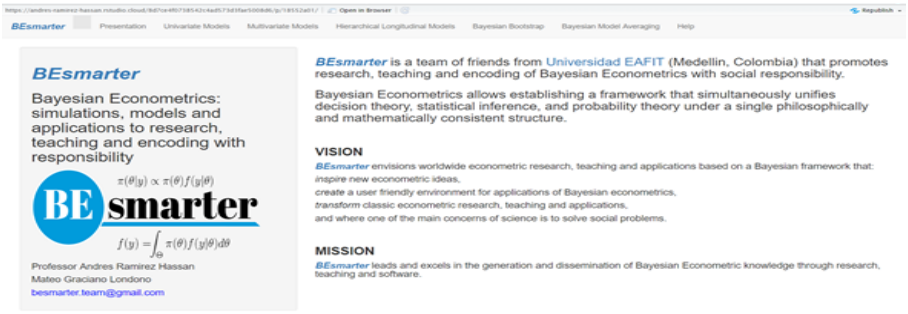

FIGURE 6.1
Display of our graphical user interface.

operating system. Our GUI, which we call BEsmarter,[1] is freely available at **https://github.com/besmarter/BSTApp**; so users have access to all our code and datasets.

6.2 Univariate models

Simulated and applied datasets are in the folders *DataSim* (see Table 15.2 for details), and *DataApp* (see Table 15.3 for details) of our *GitHub* repository. The former folder also includes the files that were used to simulate different processes, so, the population parameters are available, and as a consequence these files can be used as a pedagogical tool to show some statistical properties of the inferential frameworks available in our GUI. The latter folder contains the datasets used in the applications of this second part of the book. Users should use these datasets as templates to the structure of their own datasets. Simply type **shiny::runGitHub("besmarter/BSTApp", launch.browser=T)** in the **R** package console or any **R** code editor to run our GUI.[2]

After this, users can see a new window where a presentation of our research team is displayed. In addition, the top panel in Figure 6.1 shows the class of models that can be estimated in our GUI.

The selection indicates univariate models in that the radio button on the left hand side shows the specific models inside this generic class. In particu-

[1] Bayesian econometrics: Simulations, models and applications to research, teaching and encoding with responsibility.

[2] We strongly recommend to type this directly, rather than copy and paste. This is due to an issue with the quotation mark.

FIGURE 6.2
Univariate models: Specification.

lar, users can see that the normal model is selected from inside the class of univariate models (see Figure 6.2).

Then, the right hand side panel displays a widget to upload the input dataset, which should be a *csv* file with headers in the first row. Users also should select the kind of separator used in the input file: comma, semicolon, or tab (use the folders **DataSim** and **DataApp** for the input file templates). Once users upload the dataset, they can see a data preview. Range sliders help to set the number of iterations of the Markov chain Monte Carlo algorithm, the amount of burn-in, and the thinning parameter can be selected as well (see next chapters of this second part of the book for technical details). After this, users should specify the equation. This can be done with the formula builder, where users can select the dependent variable, and the independent variables, and then click on the *Build formula* tab. Users can see in the *Main Equation* space the formula expressed in the format used by **R** software (see Main equation box in Figure 6.2, $y \sim x1 + x2 + x3$). Users can modify this if necessary, for instance, including higher order or interaction terms, other transformations are also allowed. This is done directly in the *Main Equation* space taking into account that this extra terms should follow formula command structure.[3] Note that the class of univariate models includes the intercept by default, except ordered probit, where the specification has to do this explicitly, that is, ordered probit models do not admit an intercept, for identification issues (see details below). Hence, users should write down specifically this fact ($y \sim x1+x2+x3-1$). Finally, users should define the hyperparameters of the prior; for instance, in the normal-inverse gamma model, these are the mean vector, covariance matrix, shape, and scale parameters (see Figure 6.3). However, users should take into account that our GUI has *non-informative* hyperparameters by default in all our modelling frameworks, so the last part is not a requirement.

[3]See **https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/formula**

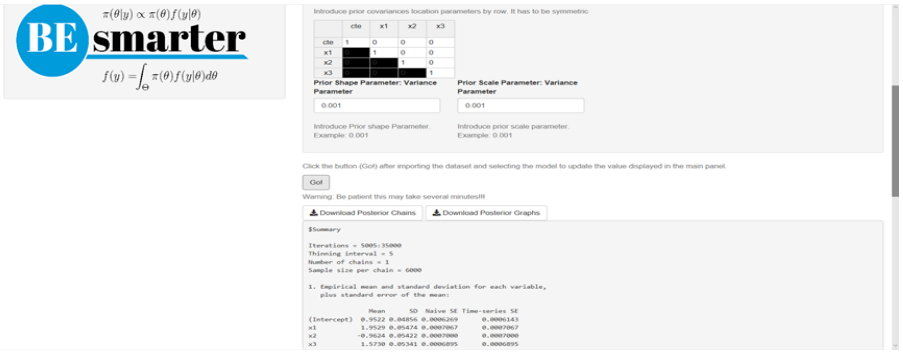

FIGURE 6.3
Univariate models: Results.

After this specification process, users should click the *Go!* button to initiate the estimation. Our GUI displays the summary statistics and convergence diagnostics after this process is finished (see Figure 6.3). There are also widgets to download posterior chains (*csv* file) and graphs (*pdf* and *eps* files). Note that the order of the coefficients in the results (summary, posterior chains, and graphs) is first for the location parameters, and then for the scale parameters.

Multinomial models (probit and logit) require a dataset file to have in the first column the dependent variable, then alternative specific regressors (for instance alternatives' prices), and finally, non-alternative regressors (for instance, income). The formula builder specifies the dependent variable, and independent variables that are alternative specific and non-alternative specific (see technical details in next chapter). Specification also requires defining the base category, number of alternatives (this is also required in ordered probit), number of alternative specific regressors, and number of non-alternative regressors (see Figure 6.4). Multinomial logit also allows defining a tuning parameter, the number of degrees of freedom in this case, for the Metropolis–Hastings algorithm (see technical details in next chapter). This is a feature in our GUI when the estimation of the models is based on the Metropolis–Hastings algorithm. The order of the coefficients in the results of these models is first the intercepts (cte_l appearing in the summary display, l-th alternative), and then the non-alternative specific regressors (NAS_{jl} appearing in the summary display, l-th alternative and j-th non-alternative regressor), and lastly, the coefficients for the alternative specific regressors (AS_j appearing in the summary display, j-th alternative specific regressor). Note that the non-alternative specific regressors associated with the base category are equal to zero (they do not appear in the results). In addition, some coefficients of the main diagonal of the covariance matrix are constant due to identification issues in multinomial and multivariate probit models.

In the case of the negative binomial model, users should set a dispersion parameter (see the negative binomial model in the next chapter). User should

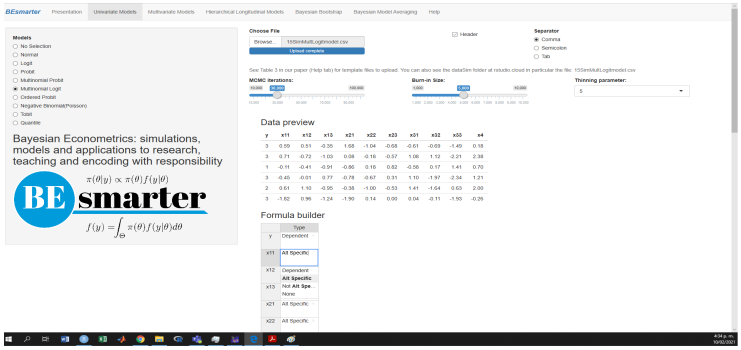

FIGURE 6.4
Univariate models: Multinomial.

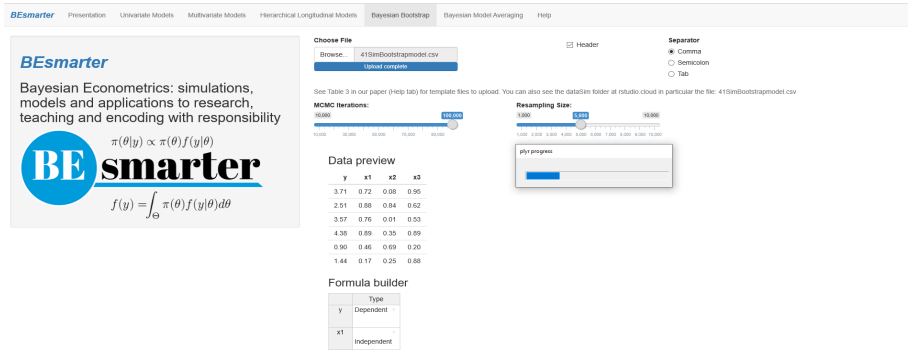

FIGURE 6.5
Univariate models: Bootstrap.

also set the censorship points and quantiles in the Tobit and quantile models, respectively.

Bayesian bootstrap only requires uploading a dataset, specifying the number of iterations of the MCMC, the resampling size, and the equation (see Figure 6.5). The input file has the same structure as the file used in the univariate normal model.

6.3 Multivariate models

Figure 6.6 displays the multivariate regression setting. In this case, the input file should have first the dependent variables, and then the regressors. If there are intercepts in each equation, there should be a column of 1's after

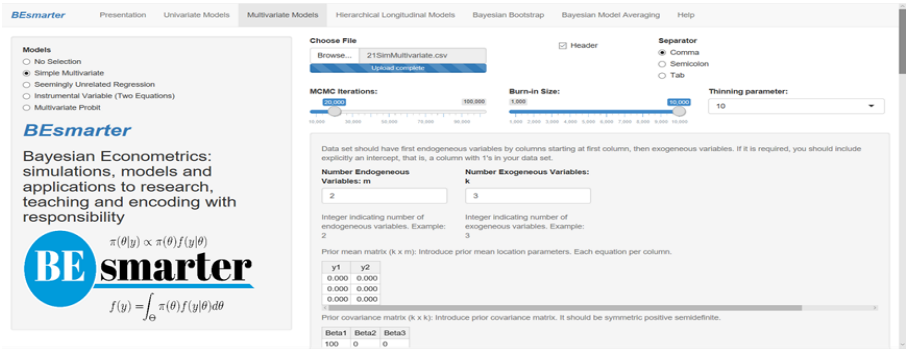

FIGURE 6.6
Multivariate models: Simple multivariate.

the dependent variables in the input file. The user also has to set the number of dependent variables, the number of regressors, if necessary include the intercept, and the values of the hyperparameters (see Figure 6.6).

The input file in seemingly unrelated regressions should have first the dependent variables, and then the regressors by equation, including the intercept in each equation if necessary (column of 1's). Users should define the number of dependent variables (equations), the number of total regressors, that is, the sum of all regressors associated with the equation (if necessary include intercepts, each intercept is an additional regressor), and the number of regressors by equation (if necessary include the intercept). Users can also set the values of the hyperparameters if there is prior information.

The results of the simple multivariate and seemingly unrelated regressions show first the posterior location parameters by equation, and then the posterior covariance matrix.

In the instrumental variable setting, users should specify the main equation and the instrumental equation. This setting includes intercepts by default. The first variable on the right hand side in the main equation has to be the variable with endogeneity issues. In the instrumental equation box, the dependent variable is the variable with endogeneity issues as a function of the instruments. Users can also specify the values of the hyperparameters if they have prior information. The input file should have the dependent variable, the endogenous regressor, the instruments, and the exogenous regressors. The results first list the posterior estimates of the endogenous regressor, then the location parameters of the auxiliary regression (instrumental equation), and the location parameters of the exogenous regressors. Last is the posterior covariance matrix.

The multivariate probit model requires an input dataset ordered by unit, for instance three choices implies repeat each unit three times. The first column has to be the identification of each unit; users should use ordered integers,

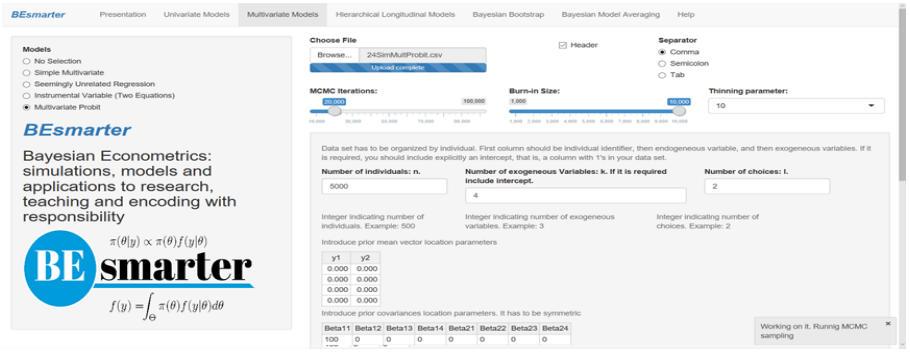

FIGURE 6.7
Multivariate models: Multivariate probit.

then the dependent variable, just one vector, composed of 0's and 1's, then the regressors, which should include a column of 1's for the intercepts. Users should set the number of units, number of regressors, and number of choices (see Figure 6.7). The results first display the posterior location parameters by equation, and then the posterior covariance matrix.

6.4 Time series model

6.5 Longitudinal (panel) models

The hierarchical longitudinal models tab allows for estimating models that account for within-subject correlation when the dependent variable is continuous (Normal), binary (Logit), or a count (Poisson).

The input files for hierarchical longitudinal models should have first the dependent variable, then the regressors and a cross sectional identifier ($i = 1, 2, \ldots, m$). It is not a requirement to have a balanced dataset: n_i can be different for each i (see Chapter 10 for technical details). Users should specify the fixed part equation and the random part equation, both in **R** format. In case of only requiring random intercepts, do not introduce anything in the latter part (see Figure 6.8). Users should also type the name of the cross sectional identifier variable. The results displayed and the posterior graphs are associated with the fixed effects and covariance matrix. However, users can download the posterior chains of all posterior estimates: fixed and random effects, and covariance matrix.

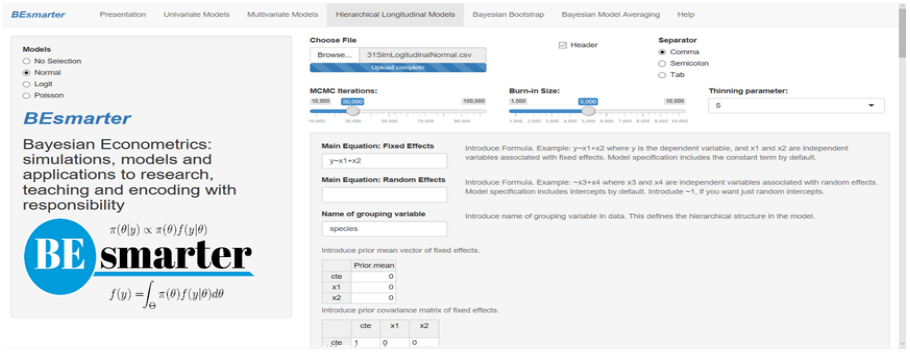

FIGURE 6.8
Hierarchical longitudinal models: Specification.

6.6 Bayesian model average

Bayesian model averaging based on a Gaussian distribution can be carried out using the Bayesian information criterion (BIC) approximation, Markov chain Monte Carlo model composition (MC3), or instrumental variables (see Figure 6.9). The former two approaches require an input dataset where the first column is the dependent variable, and then, the potentially important regressors. Users should set the band width model selection parameter (O_R) and number of iterations for BIC and MC3, respectively (see Chapter 11 for technical details). The results include the posterior inclusion probability ($p! = 0$), expected value (EV), and standard deviation (SD) of the coefficients associated with each regressor. The BIC framework also displays the most relevant models, including the number of regressors, the coefficient of determination (R^2), the BIC, and the posterior model probability. Users can download two *csv* files: *Best models* and *Descriptive statistics coefficients*. The former is a 0-1 matrix such that the columns are the regressors and the rows are the models; a 1 indicates the presence of a specific regressor in a specific model, 0 otherwise. Note that the last column of this file is the posterior model probability for each model (row). The latter file shows the posterior inclusion probabilities, expected values, and standard deviations associated with each regressor, taking into account the BMA procedure based on the best models.

Bayesian model averaging with endogeneity issues requires two input files. The first one has the dependent variable in the first column, the next columns are the regressors with endogeneity issues, and then the exogeneous regressors. The user should include a column of 1's if an intercept is required. The second input file has all the instruments. Users should also introduce the number of regressors with endogeneity issues (see Figure 6.10).

The results include the posterior inclusion probabilities and expected val-

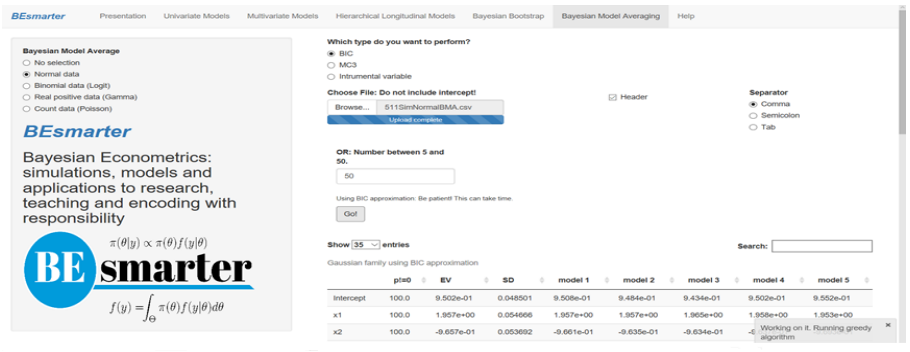

FIGURE 6.9
Bayesian model averaging: Specification and results.

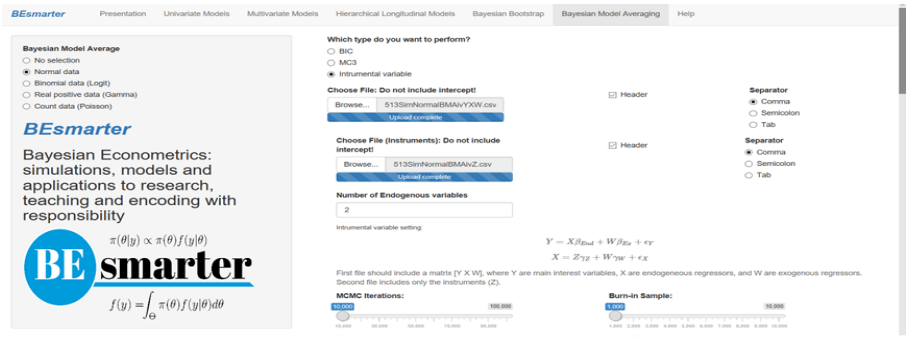

FIGURE 6.10
Bayesian model averaging: Instrumental variable specification.

ues for each regressor. The user can find the results of the main equation, and then of the auxiliary equations. Users can download *csv* files of BMA results for both the second stage (main equation) and the first stage (auxiliary equations). In addition, users can download the posterior chains of the location parameters of the main equation, $\beta_l, l = 1, 2, \ldots, dim\{\boldsymbol{\beta}\}$, the location parameters of the auxiliary equations, $\gamma_{j,i}$, $j = 1, 2, \ldots, dim\{\boldsymbol{\beta}_s\}$ where $dim\{\boldsymbol{\beta}_s\}$ is the number of regressors with endogeneity issues, $i = 1, 2, \ldots, dim\{\boldsymbol{\gamma}\}$, where $dim\{\boldsymbol{\gamma}\}$ is the number of regressors in the auxiliary regressors (exogeneous regressors + instruments), and the elements of the covariance matrix $\sigma_{j,k}$ (see Chapter 11 for technical details).

Bayesian model averaging based on BIC approximation for non-linear models, Logit, Gamma, and Poisson, requires an input dataset where the first column is the dependent variable, and the other columns are the potentially relevant regressors. Users should specify the band width model selection parameters, which are also referred to as Occam's window parameters (O_R and O_L). Our GUI displays the posterior inclusion probabilities ($p! = 0$), the expected value of the posterior coefficients (EV), and the standard deviation (SD). In addition, users can see the results associated with the models with the highest posterior model probabilities, and download *csv* files with the results of specifications of the best models, and descriptive statistics of the posterior coefficients from the BMA procedure. These files are similar to the results of the BIC approximation of the Gaussian model.

6.7 Warning

Users should also note that sometimes our GUI shuts down. In our experience, this is due to computational issues using the implicit commands that we call when estimating some models, for instance, computationally singular systems, missing values where TRUE/FALSE needed, L-BFGS-B needs finite values of "fn", NA/NaN/Inf values, or Error in backsolve. Sometimes these issues can be solved by adjusting the dataset, for instance, avoiding high levels of multicollinearity. It should also be taken into account that when warning messages are displayed in our GUI, there is a high chance that there are convergence issues of the posterior chains. So, the results are not trustworthy. Users can identify these problems by checking the console of their *RStudio* sections, where the specific folder/file where the issue happened is specified. In any case, we would appreciate your feedback to improve and enhance our GUI.

7

Univariate models

We describe how to perform Bayesian inference in some of the most common univariate models: normal-inverse gamma, logit, probit, multinomial probit and logit, ordered probit, negative binomial, tobit, quantile regression, and Bayesian bootstrap in linear models. We show the posterior distributions of the parameters and some applications. In addition, we show how to perform inference in various models using three levels of programming skills: our graphical user interface (GUI), packages from **R**, and programming the posterior distributions. The first requires no programming skills, the second requires an intermediate level, and the third demands more advanced skills. We also include mathematical and computational exercises.

We can run our GUI typing

R code. How to display our graphical user interface

```
shiny::runGitHub("besmarter/BSTApp", launch.browser = T)
```

in the **R** package console or any **R** code editor.

7.1 The Gaussian linear model

The Gaussian linear model specifies $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mu$ such that $\mu \sim N(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I_N})$ is an stochastic error, $\mathbf{X}$ is a $N \times K$ matrix of regressors, $\boldsymbol{\beta}$ is a K-dimensional vector of location coefficients, σ^2 is the variance of the model (scale parameter), $\mathbf{y}$ is a N-dimensional vector of a dependent variable, and N is the sample size. We describe this model using the conjugate family in Section 4.3, that is, $\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta}|\sigma^2) \times \pi(\sigma^2)$, and this allowed to get the posterior marginal distribution for $\boldsymbol{\beta}$ and σ^2.

We assume independent prior in this section, that is, $\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta}) \times$

$\pi(\sigma^2)$, where $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{B}_0)$ and $\sigma^2 \sim IG(\alpha_0/2, \delta_0/2)$, $\alpha_0/2$ and $\delta_0/2$ are the shape and rate parameters. This setting allows getting the posterior conditional distributions, that is, $\pi(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})$ and $\pi(\sigma^2|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$, which in turn allows to use the Gibbs sampler algorithm to perform posterior inference of $\boldsymbol{\beta}$ and σ^2.

The likelihood function in this model is

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}.$$

Then, the conditional posterior distributions are

$$\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim N(\boldsymbol{\beta}_n, \sigma^2\mathbf{B}_n),$$

and

$$\sigma^2|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X} \sim IG(\alpha_n/2, \delta_n/2),$$

where $\mathbf{B}_n = (\mathbf{B}_0^{-1} + \sigma^{-2}\mathbf{X}^\top\mathbf{X})^{-1}$, $\boldsymbol{\beta}_n = \mathbf{B}_n(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\mathbf{X}^\top\mathbf{y})$, $\alpha_n = \alpha_0 + N$ and $\delta_n = \delta_0 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ (see Exercise 1 in this chapter).[1]

Example: The market value of soccer players in Europe

Let's analyze the determinants of the market value of soccer players in Europe. In particular, we use the dataset *1ValueFootballPlayers.csv* which is in folder **DataApp** in our github repository **https://github.com/besmarter/BSTApp**. This dataset was used by [83] to finding the determinants of high performance soccer players in the five most important national leagues in Europe.

The specification of the model is

$$\begin{aligned}\log(\text{Value}_i) = \beta_1 + \beta_2\text{Perf}_i + \beta_3\text{Age}_i + \beta_4\text{Age}_i^2 + \beta_5\text{NatTeam}_i \\ + \beta_6\text{Goals}_i + \beta_7\text{Exp}_i + \beta_8\text{Exp}_i^2 + \mu_i,\end{aligned}$$

where *Value* is the market value in Euros (2017), *Perf* is a measure of performance, *Age* is the players' age in years, *NatTem* is an indicator variable that takes the value of 1 if the player has been on the national team, *Goals* is the number of goals scored by the player during his career, and *Exp* is his experience in years.

We assume that the dependent variable distributes normal, then we use a normal-inverse gamma model using vague conjugate priors where $\mathbf{B}_0 = 1000\mathbf{I}_8$, $\boldsymbol{\beta}_0 = \mathbf{0}_8$, $\alpha_0 = 0.001$ and $\delta_0 = 0.001$. We perform a Gibbs sampler with 5,000 MCMC iterations plus a burn-in equal to 5,000, and a thinning parameter equal to 1.

Once our GUI is displayed (see beginning of this chapter), we should follow Algorithm A1 to run linear Gaussian models in our GUI (see Chapter 6 for details):

[1] This model can be extended to consider heteroskedasticity such that $y_i \sim N(\mathbf{x}_i^\top\boldsymbol{\beta}, \sigma^2/\tau_i)$, where $\tau_i \sim G(v/2, v/2)$. See exercise 2 for details.

Algorithm A1 Linear Gaussian model

1: Select *Univariate Models* on the top panel
2: Select *Normal* model using the left radio button
3: Upload the dataset selecting first if there is header in the file, and the kind of separator in the *csv* file of the dataset (comma, semicolon, or tab). Then, use the *Browse* button under the **Choose File** legend. You should see a preview of the dataset
4: Select MCMC iterations, burn-in and thinning parameters using the *Range sliders*
5: Select dependent and independent variables using the *Formula builder* table
6: Click the *Build formula* button to generate the formula in **R** syntax. You can modify the formula in the **Main equation** box using valid arguments of the *formula* command structure in **R**
7: Set the hyperparameters: mean vector, covariance matrix, shape and scale parameters. This step is not necessary as by default our GUI uses non-informative priors
8: Click the *Go!* button
9: Analyze results
10: Download posterior chains and diagnostic plots using the *Download Posterior Chains* and *Download Posterior Graphs* buttons

We can see in the next **R** codes how to perform the linear Gaussian model using the command *MCMCregress* of the *MCMCpack* package, and programming the Gibbs sampler ourselves. We should get similar results using the three approaches: GUI, package and our function. In fact, our GUI relies on the *MCMCregress* command. For instance, the value of a top soccer player in Europe increases 134% ($\exp(0.85) - 1$)) on average when he has played in the national team, the credible interval at 95% is (86%, 197%).

R code. The value of soccer players, using R packages

```
rm(list = ls())
set.seed(010101)
############################ Linear regression: Value of soccer players ############################
Data <- read.csv("DataApplications/1ValueFootballPlayers.csv", sep = ",", header = TRUE, fileEncoding = "latin1")
attach(Data)
y <- log(Value)
# Value: Market value in Euros (2017) of soccer players
# Regressors quantity including intercept
X <- cbind(1, Perf, Age, Age2, NatTeam, Goals, Exp, Exp2)
# Perf: Performance. Perf2: Performance squared. Age: Age; Age: Age squared.
# NatTeam: Indicator of national team. Goals: Scored goals. Goals2: Scored goals squared
# Exp: Years of experience. Exp2: Years of experience squared. Assists: Number of assists
k <- dim(X)[2]
N <- dim(X)[1]
# Hyperparameters
d0 <- 0.001/2
a0 <- 0.001/2
b0 <- rep(0, k)
c0 <- 1000
B0 <- c0*diag(k)
B0i <- solve(B0)
# MCMC parameters
mcmc <- 5000
burnin <- 5000
tot <- mcmc + burnin
thin <- 1
# Posterior distributions using packages: MCMCpack sets the model in terms of the precision matrix
posterior  <- MCMCpack::MCMCregress(y~X-1, b0=b0, B0 = B0i, c0 = a0, d0 = d0, burnin = burnin, mcmc = mcmc, thin = thin)
summary(coda::mcmc(posterior))
Iterations = 1:5000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 5000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
           Mean        SD  Naive SE Time-series SE
X          3.695499 2.228060 3.151e-02      3.151e-02
XPerf      0.035445 0.004299 6.079e-05      6.079e-05
XAge       0.778410 0.181362 2.565e-03      2.565e-03
XAge2     -0.016617 0.003380 4.781e-05      4.781e-05
XNatTeam   0.850362 0.116861 1.653e-03      1.689e-03
XGoals     0.009097 0.001603 2.266e-05      2.266e-05
XExp       0.206208 0.062713 8.869e-04      8.428e-04
XExp2     -0.006992 0.002718 3.844e-05      3.719e-05
sigma2     0.969590 0.076091 1.076e-03      1.076e-03
```

R. code. The value of soccer players, programming our Gibbs sampler

```
# Posterior distributions programming the Gibbs sampling
# Auxiliary parameters
XtX <- t(X)%*%X
bhat <- solve(XtX)%*%t(X)%*%y
an <- a0 + N
# Gibbs sampling functions
PostSig2 <- function(Beta){
  dn <- d0 + t(y - X%*%Beta)%*%(y - X%*%Beta)
  sig2 <- invgamma::rinvgamma(1, shape = an/2, rate = dn/2)
  return(sig2)
}
PostBeta <- function(sig2){
  Bn <- solve(B0i + sig2^(-1)*XtX)
  bn <- Bn%*%(B0i%*%b0 + sig2^(-1)*XtX%*%bhat)
  Beta <- MASS::mvrnorm(1, bn, Bn)
  return(Beta)
}
PostBetas <- matrix(0, mcmc+burnin, k)
PostSigma2 <- rep(0, mcmc+burnin)
Beta <- rep(0, k)
for(s in 1:tot){
  sig2 <- PostSig2(Beta = Beta)
  PostSigma2[s] <- sig2
  Beta <- PostBeta(sig2 = sig2)
  PostBetas[s,] <- Beta
}
keep <- seq((burnin+1), tot, thin)
PosteriorBetas <- PostBetas[keep,]
colnames(PosteriorBetas) <- c("Intercept", "Perf", "Age", "
    Age2", "NatTeam", "Goals", "Exp", "Exp2")
summary(coda::mcmc(PosteriorBetas))
Iterations = 1:5000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 5000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
          Mean       SD  Naive SE Time-series SE
Intercept  3.663230 2.194363 3.103e-02      3.103e-02
Perf       0.035361 0.004315 6.102e-05      6.102e-05
Age        0.780374 0.178530 2.525e-03      2.525e-03
Age2      -0.016641 0.003332 4.713e-05      4.713e-05
NatTeam    0.850094 0.119093 1.684e-03      1.684e-03
Goals      0.009164 0.001605 2.270e-05      2.270e-05
Exp        0.205965 0.062985 8.907e-04      8.596e-04
Exp2      -0.007006 0.002731 3.862e-05      3.701e-05
PosteriorSigma2 <- PostSigma2[keep]
summary(coda::mcmc(PosteriorSigma2))
Iterations = 1:5000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 5000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
Mean             SD       Naive SE Time-series SE
0.973309       0.077316       0.001093       0.001116
```

7.2 The logit model

In the logit model the dependent variable is binary, $Y_i = \{1, 0\}$, then it follows a Bernoulli distribution, $Y_i \overset{ind}{\sim} B(\pi_i)$, that is, $p(Y_i = 1) = \pi_i$, such that $\pi_i = \frac{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}$.

The likelihood function of the logit model is

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}) &= \prod_{i=1}^{N} \pi_i^{y_i} (1-\pi_i)^{1-y_i} \\ &= \prod_{i=1}^{N} \left(\frac{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} \right)^{y_i} \left(\frac{1}{1+\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} \right)^{1-y_i}. \end{aligned}$$

We can specify a Normal distribution as prior, $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{B}_0)$. Then, the posterior distribution is

$$\begin{aligned} \pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) \propto & \prod_{i=1}^{N} \left(\frac{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} \right)^{y_i} \left(\frac{1}{1+\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} \right)^{1-y_i} \\ & \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{B}_\mathbf{0}^{-\mathbf{1}} (\boldsymbol{\beta} - \boldsymbol{\beta}_\mathbf{0}) \right\}. \end{aligned}$$

The logit model does not have a standard posterior distribution. Then, a random walk Metropolis–Hastings algorithm can be used to obtain draws from the posterior distribution. A potential proposal is a multivariate Normal centered at the current value, with covariance matrix $\tau^2(\mathbf{B}_0^{-1} + \widehat{\boldsymbol{\Sigma}}^{-1})^{-1}$, where $\tau > 0$ is a tuning parameter and $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix from the maximum likelihood estimation [59].[2]

Observe that $\log(p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X})) = \sum_{i=1}^{N} y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta}))$. We can use this expression when calculating the acceptance parameter in the computational implementation of the Metropolis-Hastings algorithm. In particular, the acceptance parameter is

$$\alpha = \min\left\{1, \exp(\log(p(\mathbf{y}|\boldsymbol{\beta}^c, \mathbf{X})) + \log(\pi(\boldsymbol{\beta}^c)) - (\log(p(\mathbf{y}|\boldsymbol{\beta}^{(s-1)}, \mathbf{X})) + \log(\pi(\boldsymbol{\beta}^{(s-1)})))\right\},$$

where $\boldsymbol{\beta}^c$ and $\boldsymbol{\beta}^{(s-1)}$ are the draws from the proposal distribution and the previous iteration of the Markov chain, respectively.[3]

Example: Simulation exercise

[2] Tuning parameters should be set in a way such that one obtains reasonable diagnostic criteria and acceptation rates.

[3] Formulating the acceptance rate using log helps to mitigate computational problems.

Let's do a simulation exercise to check the performance of the algorithm. Set $\boldsymbol{\beta} = \begin{bmatrix} 0.5 & 0.8 & -1.2 \end{bmatrix}^\top$, $x_{ik} \sim N(0,1)$, $k = 2, 3$ and $i = 1, 2, \ldots, 10000$.

We set as hyperparameters $\boldsymbol{\beta}_0 = [0\ 0\ 0]^\top$ and $\mathbf{B}_0 = 1000\mathbf{I}_3$. The tune parameter for the Metropolis-Hastings algorithm is equal to 1.

Once our GUI is displayed (see beginning of this chapter), we should follow Algorithm A2 to run logit models in our GUI (see Chapter 6 for details):

Algorithm A2 Logit model

1: Select *Univariate Models* on the top panel
2: Select *Logit* model using the left radio button
3: Upload the dataset selecting first if there is header in the file, and the kind of separator in the *csv* file of the dataset (comma, semicolon, or tab). Then, use the *Browse* button under the **Choose File** legend. You should see a preview of the dataset
4: Select MCMC iterations, burn-in and thinning parameters using the *Range sliders*
5: Select dependent and independent variables using the *Formula builder* table
6: Click the *Build formula* button to generate the formula in **R** syntax. You can modify the formula in the **Main equation** box using valid arguments of the *formula* command structure in **R**
7: Set the hyperparameters: mean vector and covariance matrix. This step is not necessary as by default our GUI uses non-informative priors
8: Select the tuning parameter for the Metropolis-Hastings algorithm
9: Click the *Go!* button
10: Analyze results
11: Download posterior chains and diagnostic plots using the *Download Posterior Chains* and *Download Posterior Graphs* buttons

We can see in the next **R** codes how to perform the logit model using the command *MCMClogit* of the *MCMCpack* package, and programming the Metropolis-Hastings algorithm ourselves.

We should get similar results using the three approaches: GUI, package and our function. Our GUI relies on the *MCMClogit* command. In particular, we obtain an acceptance rate of 0.46, and the diagnostics suggest that the posterior chains behave well. In general, the 95% credible intervals encompass the population values, and the mean and median are very close to these values.

R. code. Simulation of the logit model estimation using R packages

```
############################ Logit: Simulation ############################
# Simulate data
rm(list = ls())
set.seed(010101)
N <- 10000 # Sample size
B <- c(0.5, 0.8, -1.2) # Population location parameters
x2 <- rnorm(N) # Regressor
x3 <- rnorm(N) # Regressor
X <- cbind(1, x2, x3) # Regressors
XB <- X%*%B
PY <- exp(XB)/(1 + exp(XB)) # Probability of Y = 1
Y <- rbinom(N, 1, PY) # Draw Y's
table(Y) # Frequency
# write.csv(cbind(Y, x2, x3), file = "DataSimulations/LogitSim.csv") # Export data
# MCMC parameters
iter <- 5000; burnin <- 1000; thin <- 5; tune <- 1
# Hyperparameters
K <- dim(X)[2]
b0 <- rep(0, K)
c0 <- 1000
B0 <- c0*diag(K)
B0i <- solve(B0)
# Posterior distributions using packages: MCMCpack sets the model in terms of the precision matrix
RegLog <- MCMCpack::MCMClogit(Y~X-1, mcmc = iter, burnin = burnin, thin = thin, b0 = b0, B0 = B0i, tune = tune)
summary(RegLog)
Iterations = 1001:5996
Thinning interval = 5
Number of chains = 1
Sample size per chain = 1000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
      Mean       SD  Naive SE Time-series SE
X    0.4896 0.02550 0.0008064       0.001246
Xx2  0.8330 0.02730 0.0008632       0.001406
Xx3 -1.2104 0.03049 0.0009643       0.001536
2. Quantiles for each variable:
      2.5%     25%     50%     75%   97.5%
X    0.4424  0.4728  0.4894  0.5072  0.5405
Xx2  0.7787  0.8159  0.8327  0.8505  0.8852
Xx3 -1.2758 -1.2296 -1.2088 -1.1902 -1.1513
```

R. code. Simulation of the logit model estimation programming our M-H algorithm

```
# Posterior distributions programming the Metropolis-Hastings algorithm
MHfunc <- function(y, X, b0 = rep(0, dim(X)[2] + 1), B0 = 1000*diag(dim(X)[2] + 1), tau = 1, iter = 6000, burnin = 1000, thin = 5){
  Xm <- cbind(1, X) # Regressors
  K <- dim(Xm)[2] # Number of location parameters
  BETAS <- matrix(0, iter + burnin, K) # Space for posterior chains
  Reg <- glm(y ~ Xm - 1, family = binomial(link = "logit")) # Maximum likelihood estimation
  BETA <- Reg$coefficients # Maximum likelihood parameter estimates
  tot <- iter + burnin # Total iterations M-H algorithm
  COV <- vcov(Reg) # Maximum likelihood covariance matrix
  COVt <- tau^2*solve(solve(B0) + solve(COV)) # Covariance matrix for the proposal distribution
  Accep <- rep(0, tot) # Space for calculating the acceptance rate
  # Create progress bar in case that you want to see iterations progress
  pb <- winProgressBar(title = "progress bar", min = 0,
  max = tot, width = 300)
  for(it in 1:tot){
    BETAc <- BETA + MASS::mvrnorm(n = 1, mu = rep(0, K), Sigma = COVt) # Candidate location parameter
    likecand <- sum((Xm%*%BETAc) * Y - apply(Xm%*%BETAc, 1, function(x) log(1 + exp(x)))) # Log likelihood for the candidate
    likepast <- sum((Xm%*%BETA) * Y - apply((Xm%*%BETA), 1, function(x) log(1 + exp(x)))) # Log likelihood for the actual draw
    priorcand <- (-1/2)*crossprod((BETAc - b0), solve(B0))%*%(BETAc - b0) # Log prior for candidate
    priorpast <- (-1/2)*crossprod((BETA - b0), solve(B0))%*%(BETA - b0) # Log prior for actual draw
    alpha <- min(1, exp((likecand - priorcand) - (likepast - priorpast))) #Probability of selecting candidate
    u <- runif(1) # Decision rule for selecting candidate
    if(u < alpha){
      BETA <- BETAc # Changing reference for candidate if selected
      Accep[it] <- 1 # Indicator if the candidate is accepted
    }
    BETAS[it, ] <- BETA # Saving draws
    setWinProgressBar(pb, it, title=paste( round(it/tot*100, 0),
    "% done"))
  }
  close(pb)
  keep <- seq(burnin, tot, thin)
  return(list(Bs = BETAS[keep[-1], ], AceptRate = mean(Accep[keep[-1]])))
}
```

R. code. Simulation of the logit model programming our M-H algorithm, results

```
Posterior <- MHfunc(y = Y, X = cbind(x2, x3), iter = iter,
    burnin = burnin, thin = thin) # Running our M-H function
     changing some default parameters.
paste("Acceptance rate equal to", round(Posterior$AceptRate,
    2), sep = " ")
"Acceptance rate equal to 0.46"
PostPar <- coda::mcmc(Posterior$Bs)
# Names
colnames(PostPar) <- c("Cte", "x1", "x2")
# Summary posterior draws
summary(PostPar)
Iterations = 1:1000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
Mean       SD  Naive SE Time-series SE
Cte  0.4893 0.02427 0.0007674       0.001223
x1   0.8309 0.02699 0.0008536       0.001440
x2  -1.2107 0.02943 0.0009308       0.001423
2. Quantiles for each variable:
    2.5%     25%     50%     75%   97.5%
Cte  0.4431  0.4721  0.4899  0.5059  0.5344
x1   0.7817  0.8123  0.8305  0.8505  0.8833
x2  -1.2665 -1.2309 -1.2107 -1.1911 -1.1538
# Trace and density plots
plot(PostPar)
# Autocorrelation plots
coda::autocorr.plot(PostPar)
# Convergence diagnostics
coda::geweke.diag(PostPar)
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5
Cte     x1     x2
-0.975 -3.112  1.326
coda::raftery.diag(PostPar,q=0.5,r=0.05,s = 0.95)
Quantile (q) = 0.5
Accuracy (r) = +/- 0.05
Probability (s) = 0.95
Burn-in  Total Lower bound  Dependence
(M)      (N)   (Nmin)       factor (I)
Cte 6        731   385          1.90
x1  6        703   385          1.83
x2  6        725   385          1.88
coda::heidel.diag(PostPar)
Stationarity start     p-value
test         iteration
Cte passed         1       0.4436
x1  passed       101       0.3470
x2  passed         1       0.0872
Halfwidth Mean   Halfwidth
test
Cte passed     0.489 0.00240
x1  passed     0.832 0.00268
x2  passed    -1.211 0.00279
```

7.3 The probit model

The probit model also has as dependent variable a binary outcome. In this case, there is a latent variable (y_i^*, unobserved) that defines the structure of the estimation problem. In particular,

$$Y_i = \begin{Bmatrix} 0, & Y_i^* \leq 0 \\ 1, & Y_i^* > 0 \end{Bmatrix},$$

such that $Y_i^* = \mathbf{x_i}^\top \boldsymbol{\beta} + \mu_\mathbf{i}$, $\mu_i \overset{i.i.d.}{\sim} N(0,1)$.[4] This implies $P(Y_i = 1) = \pi_i = \Phi(\mathbf{x_i}^\top \boldsymbol{\beta})$.

[1] implemented data augmentation [87] to apply a Gibbs sampling algorithm in this model. Augmenting this model with Y_i^*, we can have the likelihood contribution from observation i, $p(y_i|y_i^*) = \mathbb{1}_{y_i=0}\mathbb{1}_{y_i^*\leq 0} + \mathbb{1}_{y_i=1}\mathbb{1}_{y_i^*>0}$, where $\mathbb{1}_A$ is an indicator function that takes the value of 1 when condition A is satisfied.

The posterior distribution is $\pi(\boldsymbol{\beta}, \boldsymbol{y^*}|\boldsymbol{y}, \boldsymbol{X}) \propto \prod_{i=1}^N \left[\mathbb{1}_{y_i=0}\mathbb{1}_{y_i^*\leq 0} + \mathbb{1}_{y_i=1}\mathbb{1}_{y_i^*>0}\right] \times N_N(\boldsymbol{y}^*|\boldsymbol{X\beta}, \boldsymbol{I}_n) \times N_K(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \boldsymbol{B}_0)$ when taking a Gaussian distribution as prior $\boldsymbol{\beta} \sim N_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$. This implies

$$y_i^*|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X} \sim \begin{Bmatrix} TN_{(-\infty,0]}(\mathbf{x_i}^\top \boldsymbol{\beta}, \mathbf{1}) & , \ \mathbf{y_i = 0} \\ TN_{(0,\infty)}(\mathbf{x_i}^\top \boldsymbol{\beta}, \mathbf{1}) & , \ \mathbf{y_i = 1} \end{Bmatrix},^5$$

$$\boldsymbol{\beta}|\boldsymbol{y}^*, \boldsymbol{X} \sim N(\boldsymbol{\beta}_n, \boldsymbol{B}_n),$$

where $\boldsymbol{B}_n = (\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{X})^{-1}$, and $\boldsymbol{\beta}_n = \boldsymbol{B}_n(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{y}^*)$.

Example: Determinants of hospitalization

We use the dataset named **2HealthMed.csv**, which is in folder **DataApp** in our github repository **(https://github.com/besmarter/BSTApp** and was used by [73]. Our dependent variable is a binary indicator with a value equal to 1 if an individual was hospitalized in 2007, and 0 otherwise.

The specification of the model is

$$\begin{aligned} \text{Hosp}_i = \beta_1 + \beta_2 \text{SHI}_i + \beta_3 \text{Female}_i + \beta_4 \text{Age}_i + \beta_5 \text{Age}_i^2 + \beta_6 \text{Est2}_i + \beta_7 \text{Est3}_i \\ + \beta_8 \text{Fair}_i + \beta_9 \text{Good}_i + \beta_{10} \text{Excellent}_i, \end{aligned}$$

where *SHI* is a binary variable equal to 1 if the individual is in a subsidized

[4] The variance in this model is set to 1 due to identification restrictions. Observe that $P(Y_i = 1|\mathbf{x_i}) = \mathbf{P}(\mathbf{Y_i^*} > \mathbf{0}|\mathbf{x_i}) = \mathbf{P}(\mathbf{x_i}^\top \boldsymbol{\beta} + \mu_\mathbf{i} > \mathbf{0}|\mathbf{x_i}) = \mathbf{P}(\mu_\mathbf{i} > -\mathbf{x_i}^\top \boldsymbol{\beta}|\mathbf{x_i}) = \mathbf{P}(\mathbf{c} \times \mu_\mathbf{i} > -\mathbf{c} \times \mathbf{x_i}^\top \boldsymbol{\beta}|\mathbf{x_i})$ $\forall c > 0$. Multiplying for a positive constant does not affect the probability of $Y_i = 1$.

[5] TN denotes a truncated normal density.

health care program and 0 otherwise, *Female* is an indicator of gender, *Age* in years, *Est2* and *Est3* are indicators of socioeconomic status, the reference is *Est1*, which is the lowest, and self perception of health status where *bad* is the reference.

Let's set $\boldsymbol{\beta}_0 = \mathbf{0}_{10}$, $\mathbf{B}_0 = \mathbf{I}_{10}$, iterations, burn-in and thinning parameters equal to 10000, 1000 and 1, respectively. We can use the Algorithm A1 to run the probit model in our GUI. We should select *Probit* model in stage 2. Our GUI relies in the command *rbprobitGibbs* from the package *bayesm* to perform inference in the Probit model. The following **R** code shows how to run this example using the command *rbprobitGibbs*. We asked to program a Gibbs sampler algorithm to perform inference in the probit model in the exercises.

We find evidence that gender and self-perceived health status affect the probability of hospitalization. Women have a higher probability of being hospitalized than men, and a better perception of health status decreases this probability.

R. code. Determinants of hospitalization

```
mydata <- read.csv("DataApplications/2HealthMed.csv", header = T, sep = ",")
attach(mydata)
str(mydata)
K <- 10 # Number of regressors
b0 <- rep(0, K) # Prio mean
B0i <- diag(K) # Prior precision (inverse of covariance)
Prior <- list(betabar = b0, A = B0i) # Prior list
y <- Hosp # Dependent variables
X <- cbind(1, SHI, Female, Age, Age2, Est2, Est3, Fair, Good, Excellent) # Regressors
Data <- list(y = y, X = X) # Data list
Mcmc <- list(R = 10000, keep = 1, nprint = 0) # MCMC parameters
RegProb <- bayesm::rbprobitGibbs(Data = Data, Prior = Prior, Mcmc = Mcmc) # Inference using bayesm package
PostPar <- coda::mcmc(RegProb$betadraw) # Posterior draws
colnames(PostPar) <- c("Cte", "SHI", "Female", "Age", "Age2", "Est2", "Est3", "Fair", "Good", "Excellent") # Names
summary(PostPar) # Posterior summary
Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000
2. Quantiles for each variable:
          2.5%         25%         50%         75%       97.5%
Cte       -1.22e+00 -1.03e+00 -9.43e-01 -8.50e-01 -0.671744
SHI       -1.24e-01 -4.63e-02 -6.30e-03  3.26e-02  0.104703
Female     2.80e-02  9.65e-02  1.28e-01  1.60e-01  0.223123
Age       -7.55e-03 -2.50e-03  1.25e-04  2.80e-03  0.007646
Age2      -4.98e-05  9.05e-06  4.02e-05  7.07e-05  0.000128
Est2      -1.89e-01 -1.23e-01 -8.84e-02 -5.32e-02  0.012714
Est3      -2.13e-01 -1.03e-01 -4.73e-02  1.01e-02  0.109527
Fair      -7.09e-01 -5.69e-01 -4.93e-01 -4.16e-01 -0.269494
Good      -1.42e+00 -1.28e+00 -1.20e+00 -1.12e+00 -0.982533
Excellent -1.33e+00 -1.15e+00 -1.06e+00 -9.74e-01 -0.795881
```

7.4 The multinomial probit model

The multinomial probit model is used to model the choice of the l-th alternative over a set L mutually exclusive options. We observe

$$y_{il} = \begin{Bmatrix} 1, & y_{il}^* \geq \max\{\boldsymbol{y}_i^*\} \\ 0, & \text{otherwise} \end{Bmatrix},$$

such that $\boldsymbol{y}_i^* = \boldsymbol{X}_i\boldsymbol{\delta} + \boldsymbol{\mu}_i$, $\boldsymbol{\mu}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{y}_i^*$ is an unobserved latent L dimensional vector, $\boldsymbol{X}_i = \left[(1\ \boldsymbol{c}_i^\top) \otimes \boldsymbol{I}_L\ \boldsymbol{A}_i\right]$ is an $L \times j$ matrix of regressors for each alternative, $l = 1, 2, \ldots, L$, $j = L \times (1 + dim\{\boldsymbol{c}_i\}) + a$, $\boldsymbol{c}_i$ is a vector of the individuals' specific characteristics, $\boldsymbol{A}_i$ is an $L \times a$ matrix of alternative-varying regressors, a is the number of alternative-varying regressors, and $\boldsymbol{\delta}$ is a j dimensional vector of parameters.

We take into account simultaneously the alternative-varying regressors (alternative attributes) and alternative-invariant regressors (individual characteristics).[6] $\boldsymbol{y}_i^*$ can be stacked up into a multiple regression with correlated stochastic errors, $\boldsymbol{y}^* = \boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{\mu}$, where $\boldsymbol{y}^* = \left[\boldsymbol{y}_1^{*\top}, \boldsymbol{y}_2^{*\top}, \ldots, \boldsymbol{y}_N^{*\top}\right]$,$\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_N^\top\right]^\top$, and $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top, \ldots, \boldsymbol{\mu}_N^\top\right]^\top$.

Following the practice of expressing y_{il}^* relative to y_{iL}^* by letting $\boldsymbol{w}_i = [w_{i1}, w_{i2}, \ldots, w_{iL-1}]^\top$, $w_{il} = y_{il}^* - y_{iL}^*$, we can write $\boldsymbol{w}_i = \boldsymbol{R}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, $\boldsymbol{\epsilon_i} \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$, where $\boldsymbol{R}_i = \left[(1\ \boldsymbol{c}_i^\top) \otimes \boldsymbol{I}_{L-1}\ \boldsymbol{\Delta}\boldsymbol{A}_i\right]$ is an $L-1 \times k$ matrix where $\Delta\boldsymbol{A}_i = \boldsymbol{A}_{li} - \boldsymbol{A}_{Li}$, $l = 1, 2, \ldots, L-1$, that is, the last row of $\boldsymbol{A}_i$ is subtracted from each row of $\boldsymbol{A}_i$, and $\boldsymbol{\beta}$ is a k dimensional vector, $k = (L-1) \times (1 + dim\{\boldsymbol{c}_i\}) + a$.

Observe that $\boldsymbol{\beta}$ contains the same last a elements as $\boldsymbol{\delta}$, that is, alternative specific attributes coefficients, but the first $(L-1)\times(1+dim\{\boldsymbol{c}_i\})$-th elements are $\delta_{jl} - \delta_{jL}$, $j = 1 + dim\{\boldsymbol{c}_i\}$, $l = 1, 2, \ldots, L-1$, that is, the difference between the coefficients of each qualitative response and the L-th alternative for the individuals' characteristics.This makes it difficult to interpret the multinomial probit coefficients.

Note that in multinomial models, for each alternative specific attribute, it is only required to estimate one coefficient for all alternatives, whereas for individuals' characteristics (non-alternative specific regressors), it is necessary to estimate $L-1$ coefficients (the coefficient of the base alternative is set equal to 0).

The likelihood function in this model is $p(\boldsymbol{\beta}, \boldsymbol{\Omega}|\boldsymbol{y}, \boldsymbol{R}) = \prod_{i=1}^{N}\prod_{l=1}^{L} p_{il}^{y_{il}}$ where $p_{il} = p(y_{il}^* \geq \max(\boldsymbol{y}_i^*))$.

We assume independent priors, $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$ and $\boldsymbol{\Omega}^{-1} \sim W(\alpha_0, \boldsymbol{\Sigma}_0)$.[7] We can employ Gibbs sampling in this model because this is a standard

[6] Note that this model is not identified if $\boldsymbol{\Sigma}$ is unrestricted. The likelihood function is the same if a scalar random variable is added to each of the L latent regressions.

[7] W denotes the Wishart density.

Bayesian linear regression model when data augmentation in $\boldsymbol{w}$ is used. The posterior conditional distributions are

$$\boldsymbol{\beta}|\boldsymbol{\Omega}, \boldsymbol{w} \sim N(\boldsymbol{\beta}_n, \boldsymbol{B}_n),$$

$$\boldsymbol{\Omega}^{-1}|\boldsymbol{\beta}, \boldsymbol{w} \sim W(\alpha_n, \boldsymbol{\Sigma}_n),$$

where $\boldsymbol{B}_n = (\boldsymbol{B}_0^{-1} + \boldsymbol{X}^{*\top}\boldsymbol{X}^*)^{-1}$, $\boldsymbol{\beta}_n = \boldsymbol{B}_n(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^{*\top}\boldsymbol{w}^*)$, $\boldsymbol{\Omega}^{-1} = \boldsymbol{C}^\top\boldsymbol{C}$, $\boldsymbol{X}_i^{*\top} = \boldsymbol{C}^\top\boldsymbol{R}_i$, $\boldsymbol{w}_i^* = \boldsymbol{C}^\top\boldsymbol{w}_i$, $\boldsymbol{X}^* = \begin{bmatrix} \boldsymbol{X}_1^* \\ \boldsymbol{X}_2^* \\ \vdots \\ \boldsymbol{X}_N^* \end{bmatrix}$, $\alpha_n = \alpha_0 + N$, $\boldsymbol{\Sigma}_n = (\boldsymbol{\Sigma}_0 + \sum_{i=1}^N (\boldsymbol{w}_i - \boldsymbol{R}_i\boldsymbol{\beta})^\top(\boldsymbol{w}_i - \boldsymbol{R}_i\boldsymbol{\beta}))^{-1}$.

We can collapse the multinomial vector $\boldsymbol{y}_i$ into the indicator variable $d_i = \sum_{l=1}^{L-1} l \times \mathbb{1}_{\max(\boldsymbol{w}_l)=w_{il}}$.[8] Then the distribution of $\boldsymbol{w}_i|\boldsymbol{\beta}, \boldsymbol{\Omega}^{-1}, d_i$ is an $L-1$ dimensional Gaussian distribution truncated over the appropriate cone in $\mathcal{R}^{L-1}$. [61] propose drawing from the univariate conditional distributions $w_{il}|\boldsymbol{w}_{i,-l}, \boldsymbol{\beta}, \boldsymbol{\Omega}^{-1}, d_i \sim TN_{I_{il}}(m_{il}, \tau_{ll}^2)$, where

$$I_{il} = \begin{Bmatrix} w_{il} > \max(\boldsymbol{w}_{i,-l}, 0), & d_i = l \\ w_{il} < \max(\boldsymbol{w}_{i,-l}, 0), & d_i \neq l \end{Bmatrix},$$

and permuting the columns and rows of $\boldsymbol{\Omega}^{-1}$ so that the l-th column and row is the last,

$$\boldsymbol{\Omega}^{-1} = \begin{bmatrix} \boldsymbol{\Omega}_{-l,-l} & \boldsymbol{\omega}_{-l,l} \\ \boldsymbol{\omega}_{l,-1} & \omega_{l,l} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Omega}_{-l,-l}^{-1} + \tau_{ll}^{-2}\boldsymbol{f}\boldsymbol{f}^\top & -\boldsymbol{f}\tau_{ll}^{-2} \\ -\tau_{ll}^{-2}\boldsymbol{f}^\top & \tau_{ll}^{-2} \end{bmatrix}$$

where $\boldsymbol{f} = \boldsymbol{\Omega}_{-l,-l}^{-1}\boldsymbol{\omega}_{-l,l}$, $\tau_{ll}^2 = \omega_{ll} - \boldsymbol{\omega}_{l,-l}\boldsymbol{\Omega}_{-l,-1}^{-1}\boldsymbol{\omega}_{-l,l}$, $m_{il} = \boldsymbol{r}_{il}^\top\boldsymbol{\beta} + \boldsymbol{f}^\top(\boldsymbol{w}_{i,-l} - \boldsymbol{R}_{i,-l}\boldsymbol{\beta})$, $\boldsymbol{w}_{i,-l}$ is an $L-2$ dimensional vector of all components of $\boldsymbol{w}_i$ excluding w_{il}, $\boldsymbol{r}_{il}$ is the l-th row of $\boldsymbol{R}_i$, $l = 1, 2, \ldots, L-1$.

The identified parameters are obtained by normalizing with respect to one of the diagonal elements $\frac{1}{\omega_{1,1}^{0.5}}\boldsymbol{\beta}$ and $\frac{1}{\omega_{1,1}}\boldsymbol{\Omega}$.[9]

Example: Choice of fishing mode

We used in this application the dataset *3Fishing.csv* from [15, p. 491]. The dependent variable is mutually exclusive alternatives regarding fishing modes (mode), where beach is equal to 1, pier is equal to 2, private boat is equal to 3, and chartered boat (baseline alternative) is equal to 4. In this model, we have

$$\boldsymbol{X}_i = \begin{Bmatrix} 1 & 0 & 0 & 0 & \text{Income}_i & 0 & 0 & 0 & \text{Price}_{i,1} & \text{Catch rate}_{i,1} \\ 0 & 1 & 0 & 0 & 0 & \text{Income}_i & 0 & 0 & \text{Price}_{i,2} & \text{Catch rate}_{i,2} \\ 0 & 0 & 1 & 0 & 0 & 0 & \text{Income}_i & 0 & \text{Price}_{i,3} & \text{Catch rate}_{i,3} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \text{Income}_i & \text{Price}_{i,4} & \text{Catch rate}_{i,4} \end{Bmatrix}.$$

[8] Observe that the identification issue in this model is due to scaling w_{il} by a positive constant does not change the value of d_i.

[9] Our GUI is based on the *bayesm* package that takes into account this identification restriction to display the outcomes of the posterior chains.

In this example chartered boat is the base category, the number of choice categories is four, there are two alternative-specific regressors (price and catch rate), and one non alternative-specific regressor (income). This setting involves the estimation of eight location parameters ($\boldsymbol{\beta}$): three intercepts, three for income, one for price, and one for catch rate. This is the order of the posterior chains in our GUI. Note that the location coefficients are set equal to 0 for the baseline category. For multinomial models, we strongly recommend using the last category as the baseline.

We also get posterior estimates for a 3×3 covariance matrix (four alternatives minus one), where the element (1,1) is equal to 1 due to identification restrictions, and elements 2 and 4 are the same, as well as 3 and 7, and 6 and 8, due to symmetry.[10] Observe that this identification restriction implies *NaN* values in [36] and [42] tests for element (1,1) of the covariance matrix, and just eight dependence factors associated with the remaining elements of the covariance matrix.

Once our GUI is displayed (see beginning of this chapter), we should follow Algorithm A3 to run multinomial probit models in our GUI (see Chapter 6 for details):

Algorithm A3 Multinomial probit models

1: Select *Univariate Models* on the top panel
2: Select *Multinomial Probit* model using the left radio button
3: Upload the dataset selecting first if there is header in the file, and the kind of separator in the *csv* file of the dataset (comma, semicolon, or tab). Then, use the *Browse* button under the **Choose File** legend. You should see a preview of the dataset
4: Select MCMC iterations, burn-in and thinning parameters using the *Range sliders*
5: Select dependent and independent variables using the *Formula builder* table
6: Select the number of the **Base Alternative**
7: Select the **Number of choice categorical alternatives**
8: Select the **Number of alternative specific variables**
9: Select the **Number of Non-alternative specific variables**
10: Click the *Build formula* button to generate the formula in **R** syntax.
11: Set the hyperparameters: mean vector, covariance matrix, scale matrix and degrees of freedom. This step is not necessary as by default our GUI uses non-informative priors
12: Click the *Go!* button
13: Analyze results
14: Download posterior chains and diagnostic plots using the *Download Posterior Chains* and *Download Posterior Graphs* buttons

[10] This is the order in the pdf, eps and csv files that can be downloaded from our GUI.

We ran 100,000 MCMC iterations plus 10,000 as burn-in with a thinning parameter equal to 5, where all priors use default values for the hyperparameters in our GUI. We found that the 95% credible intervals of the coefficient associated with income for beach and private boat alternatives are equal to (8.58e-06, 8.88e-05) and (3.36e-05, 1.45e-04). This suggests that the probability of choosing these alternatives increases compared to a chartered boat when income increases. In addition, an increase in the price or a decrease in the catch rate for specific fishing alternatives imply lower probabilities of choosing them as the 95% credible intervals are (-9.91e-03, -3.83e-03) and (1.40e-01, 4.62e-01), respectively. However, the chain diagnostics suggest there are convergence issues with the posterior draws (see exercise 5).

R. code. Choice of fishing mode, results

```
Iterations = 10005:110000
Thinning interval = 5
Number of chains = 1
Sample size per chain = 20000
Quantiles for each variable:
          2.5%         25%         50%         75%       97.5%
cte_1   -5.83e-01 -4.08e-01 -3.22e-01 -2.37e-01 -7.93e-02
cte_2   -1.93e-01 -4.14e-02  2.16e-02  7.93e-02  1.93e-01
cte_3   -8.15e-01 -5.43e-01 -4.29e-01 -3.33e-01 -1.70e-01
NAS_1_1  8.58e-06  3.61e-05  4.95e-05  6.27e-05  8.88e-05
NAS_1_2 -3.24e-05 -7.04e-06  5.52e-06  1.93e-05  5.17e-05
NAS_1_3  3.36e-05  6.38e-05  8.08e-05  9.99e-05  1.45e-04
AS_1    -9.91e-03 -7.90e-03 -6.86e-03 -5.93e-03 -3.83e-03
AS_2     1.40e-01  2.25e-01  2.72e-01  3.28e-01  4.62e-01
```

7.5 The multinomial logit model

The multinomial logit model is used to model mutually exclusive discrete outcomes or qualitative response variables. However, this model assumes the independence of irrelevant alternatives (IIA), meaning that the choice between two alternatives does not depend on a third alternative. We consider the multinomial mixed logit model (not to be confused with the random parameters logit model), which accounts for both alternative-varying regressors (conditional) and alternative-invariant regressors (multinomial) simultaneously.[11]

[11] The multinomial mixed logit model can be implemented as a conditional logit model.

In this setting there are L mutually exclusive alternatives, and the dependent variable y_{il} is equal to 1 if the lth alternative is chosen by individual i, and 0 otherwise, $l = \{1, 2, \ldots, L\}$. The likelihood function is $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{N} \prod_{l=1}^{L} p_{il}^{y_{il}}$, where the probability that individual i chooses the alternative l is given by $p_{il} := p(y_i = l|\boldsymbol{\beta}, \boldsymbol{X}) = \frac{\exp\{\boldsymbol{x}_{il}^{\top}\boldsymbol{\beta}_l\}}{\sum_{j=1}^{L}\exp\{\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_j\}}$, $\boldsymbol{y}$ and $\boldsymbol{X}$ are the vector and matrix of the dependent variable and regressors, and $\boldsymbol{\beta}$ is the vector containing all the coefficients. Remember that coefficients associated with alternative-invariant regressors are set to 0 for the baseline category, and the coefficients associated with the alternative-varying regressors are the same for all the categories. In addition, we assume $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$ as prior distribution. Thus, the posterior distribution is $\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) \times \pi(\boldsymbol{\beta})$.

As the multinomial logit model does not have a standard posterior distribution, [77] propose a "tailored" independent Metropolis–Hastings algorithm where the proposal distribution is a multivariate Student's t distribution with v degrees of freedom (tuning parameter), mean equal to the maximum likelihood estimator, and scale equal to the inverse of the Hessian matrix.

Example: Simulation exercise

Let's do a simulation exercise to check the performance of the Metropolis-Hastings algorithm to perform inference in the multinomial logit model. Assume a situation where there are three alternatives, one alternative-invariant regressor plus the intercept, and three alternative-varying regressors. The population parameters are $\boldsymbol{\beta}_1 = [1\ -2.5\ 0.5\ 0.8\ -3]$, $\boldsymbol{\beta}_2 = [1\ -3.5\ 0.5\ 0.8\ -3]$ and $\boldsymbol{\beta}_3 = [0\ 0\ 0.5\ 0.8\ -3]$, the first two elements of the vectors are associated with the intercept and the alternative-invariant regressor, and the last three elements with the alternative-varying regressors. The sample size is 1000, and all regressors are simulated from standard normal distributions.

We can deploy our GUI using the command line at the beginning of this chapter. We should follow Algorithm A4 to run multinomial logit models in our GUI (see Chapter 6 for details):

The following code in **R** shows how to implement the M-H algorithm from scratch. The first part simulates the dataset, the second part builds the loglikelihood function, and the third part implements the M-H algorithm. We use vague priors centered on zero, and covariance matrix $1000\boldsymbol{I}_7$. We observe that the posterior estimates closely match the population parameters, and all 95% credible intervals contain the population parameters.

Algorithm A4 Multinomial logit models

1: Select *Univariate Models* on the top panel
2: Select *Multinomial Logit* model using the left radio button
3: Upload the dataset selecting first if there is header in the file, and the kind of separator in the *csv* file of the dataset (comma, semicolon, or tab). Then, use the *Browse* button under the **Choose File** legend. You should see a preview of the dataset
4: Select MCMC iterations, burn-in and thinning parameters using the *Range sliders*
5: Select dependent and independent variables using the *Formula builder* table
6: Select the **Base Alternative**
7: Select the **Number of choice categorical alternatives**
8: Select the **Number of alternative specific variables**
9: Select the **Number of Non-alternative specific variables**
10: Click the *Build formula* button to generate the formula in **R** syntax.
11: Set the hyperparameters: mean vector and covariance matrix. This step is not necessary as by default our GUI uses non-informative priors
12: Select the tuning parameter for the Metropolis-Hastings algorithm, that is, the **Degrees of freedom: Multivariate Student's t distribution**
13: Click the *Go!* button
14: Analyze results
15: Download posterior chains and diagnostic plots using the *Download Posterior Chains* and *Download Posterior Graphs* buttons

R. code. Simulation of the multinomial logit model

```
remove(list = ls())
set.seed(12345)
# Simulation of data
N<-1000  # Sample Size
B<-c(0.5,0.8,-3); B1<-c(-2.5,-3.5,0); B2<-c(1,1,0)
# Alternative specific attributes of choice 1, for instance,
    price, quality and duration of choice 1
X1<-matrix(cbind(rnorm(N,0,1),rnorm(N,0,1),rnorm(N,0,1)),N,
   length(B))
# Alternative specific attributes of choice 2, for instance,
    price, quality and duration of choice 2
X2<-matrix(cbind(rnorm(N,0,1),rnorm(N,0,1),rnorm(N,0,1)),N,
   length(B))
# Alternative specific attributes of choice 3, for instance,
    price, quality and duration of choice 3
X3<-matrix(cbind(rnorm(N,0,1),rnorm(N,0,1),rnorm(N,0,1)),N,
   length(B))
X4<-matrix(rnorm(N,1,1),N,1)
V1<-B2[1]+X1%*%B+B1[1]*X4; V2<-B2[2]+X2%*%B+B1[2]*X4; V3<-B2
   [3]+X3%*%B+B1[3]*X4
suma<-exp(V1)+exp(V2)+exp(V3)
p1<-exp(V1)/suma; p2<-exp(V2)/suma; p3<-exp(V3)/suma
p<-cbind(p1,p2,p3)
y<- apply(p,1, function(x)sample(1:3, 1, prob = x, replace =
    TRUE))
y1<-y==1; y2<-y==2; y3<-y==3
```

R. code. Simulation of the multinomial logit model

```
# Log likelihood
log.L<- function(Beta){
  V1<-Beta[1]+Beta[3]*X4+X1%*%Beta[5:7]
  V2<-Beta[2]+Beta[4]*X4+X2%*%Beta[5:7]
  V3<- X3%*%Beta[5:7]
  suma<-exp(V1)+exp(V2)+exp(V3)
  p11<-exp(V1)/suma;  p22<-exp(V2)/suma;  p33<-exp(V3)/suma
  suma2<-NULL
  for(i in 1:N){
    suma1<-y1[i]*log(p11[i])+y2[i]*log(p22[i])+y3[i]*log(p33
    [i])
    suma2<-c(suma2,suma1)}
  logL<-sum(suma2)
  return(-logL)
}
# Parameters: Proposal
k <- 7
res.optim<-optim(rep(0, k), log.L, method="BFGS", hessian=
    TRUE)
MeanT <- res.optim$par
ScaleT <- as.matrix(Matrix::forceSymmetric(solve(res.optim$
    hessian))) # Force this matrix to be symmetric
# Hyperparameters: Priors
B0 <- 1000*diag(k); b0 <- rep(0, k)
MHfunction <- function(iter, tuning){
  Beta <- rep(0, k); Acept <- NULL
  BetasPost <- matrix(NA, iter, k)
  pb <- winProgressBar(title = "progress bar", min = 0, max
    = iter, width = 300)
  for(s in 1:iter){
    LogPostBeta <- -log.L(Beta) + mvtnorm::dmvnorm(Beta,
    mean = b0, sigma = B0, log = TRUE)
    BetaC <- c(LaplacesDemon::rmvt(n=1, mu = MeanT, S =
    ScaleT, df = tuning))
    LogPostBetaC <- -log.L(BetaC) + mvtnorm::dmvnorm(BetaC,
    mean = b0, sigma = B0, log = TRUE)
    alpha <- min(exp((LogPostBetaC-mvtnorm::dmvt(BetaC,
    delta = MeanT, sigma = ScaleT, df = tuning, log = TRUE))
    -(LogPostBeta-mvtnorm::dmvt(Beta, delta = MeanT, sigma =
     ScaleT, df = tuning, log = TRUE))) ,1)
    u <- runif(1)
    if(u <= alpha){
      Acepti <- 1; Beta <- BetaC
    }else{
      Acepti <- 0; Beta <- Beta
    }
    BetasPost[s, ] <- Beta; Acept <- c(Acept, Acepti)
    setWinProgressBar(pb, s, title=paste( round(s/iter*100,
    0),"% done"))
  }
  close(pb); AcepRate <- mean(Acept)
  Results <- list(AcepRate = AcepRate, BetasPost = BetasPost
    )
  return(Results)
}
```

R. code. Simulation of the multinomial logit model

```
# MCMC parameters
mcmc <- 10000; burnin <- 1000; thin <- 5; iter <- mcmc +
    burnin; keep <- seq(burnin, iter, thin); tuning <- 6 #
    Degrees of freedom
ResultsPost <- MHfunction(iter = iter, tuning = tuning)
summary(coda::mcmc(ResultsPost$BetasPost[keep[-1], ]))
Iterations = 1:2000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 2000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
      Mean       SD Naive SE Time-series SE
[1,]  0.9711 0.20162 0.004508       0.004508
[2,]  0.9742 0.20934 0.004681       0.004681
[3,] -2.4350 0.18950 0.004237       0.004137
[4,] -3.4195 0.24656 0.005513       0.005513
[5,]  0.5253 0.07396 0.001654       0.001654
[6,]  0.8061 0.08007 0.001790       0.001790
[7,] -3.0853 0.17689 0.003955       0.003955
2. Quantiles for each variable:
       2.5%     25%     50%     75%   97.5%
var1  0.5862  0.8367  0.9650  1.1017  1.3683
var2  0.5679  0.8310  0.9681  1.1151  1.3761
var3 -2.8239 -2.5607 -2.4291 -2.3050 -2.0812
var4 -3.9176 -3.5806 -3.4074 -3.2496 -2.9423
var5  0.3840  0.4761  0.5250  0.5759  0.6647
var6  0.6555  0.7494  0.8064  0.8616  0.9604
var7 -3.4476 -3.1991 -3.0777 -2.9641 -2.7500
```

7.6 Ordered probit model

The ordered probit model is used when there is a natural order in the categorical response variable. In this case, there is a latent variable $y_i^* = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \mu_i$, $\mu_i \overset{i.i.d.}{\sim} N(0,1)$ such that $y_i = l$ if and only if $\alpha_{l-1} < y_i^* \leq \alpha_l$, $l = \{1, 2, \ldots, L\}$, where $\alpha_0 = -\infty$, $\alpha_1 = 0$ and $\alpha_L = \infty$.[12] Then,

[12] Identification issues necessitate setting the variance in this model equal to 1 and $\alpha_1 = 0$. Observe that multiplying y_i^* by a positive constant or adding a constant to all of the cut-offs and subtracting the same constant from the intercept does not affect y_i.

$p(y_i = l) = \Phi(\alpha_l - \boldsymbol{x}_i^\top \boldsymbol{\beta}) - \Phi(\alpha_{l-1} - \boldsymbol{x}_i^\top \boldsymbol{\beta})$, and the likelihood function is $p(\boldsymbol{\beta}, \boldsymbol{\alpha}|\boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{N} p(y_i = l|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{X})$.

There are independent priors of this model, $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \pi(\boldsymbol{\beta}) \times \pi(\boldsymbol{\gamma})$, where $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$ and $\boldsymbol{\gamma} \sim N(\boldsymbol{\gamma}_0, \boldsymbol{\Gamma}_0)$, $\boldsymbol{\gamma} = [\gamma_2, \gamma_3, \ldots, \gamma_{L-1}]^\top$, such that $\boldsymbol{\alpha} = \left[\exp\{\gamma_2\}, \sum_{l=2}^{3} \exp\{\gamma_l\}, \ldots, \sum_{l=2}^{L-1} \exp\{\gamma_l\}\right]^\top$. The latter structure imposes the ordinal condition in the cut-offs.

This model does not have a standard conditional posterior distribution for $\boldsymbol{\gamma}$ ($\boldsymbol{\alpha}$), but it does have a standard conditional distribution for $\boldsymbol{\beta}$ once data augmentation is used. Then, we can use a Metropolis-within-Gibbs sampling algorithm. In particular, we use Gibbs sampling algorithms to draw $\boldsymbol{\beta}$ and $\boldsymbol{y}^*$,

$$\boldsymbol{\beta}|\boldsymbol{y}^*, \boldsymbol{\alpha}, \boldsymbol{X} \sim N(\boldsymbol{\beta}_n, \boldsymbol{B}_n),$$

where $\boldsymbol{B}_n = (\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{X})^{-1}$, $\boldsymbol{\beta}_n = \boldsymbol{B}_n(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{y}^*)$, and $y_i^*|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{y}, \boldsymbol{X} \sim TN_{(\alpha_{y_i-1}, \alpha_{y_i})}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, 1)$.

We use a random-walk Metropolis–Hastings algorithm for $\boldsymbol{\gamma}$ that has as proposal a Gaussian distribution with mean equal to the current value, and covariance matrix $s^2(\boldsymbol{\Gamma}_0^{-1} + \hat{\boldsymbol{\Sigma}}_\gamma^{-1})^{-1}$, where $s > 0$ is a tuning parameter, and $\hat{\boldsymbol{\Sigma}}_\gamma$ is the sample covariance matrix associated with γ from the maximum likelihood estimation.

Example: Determinants of preventive health care visits

We used the file named *2HealthMed.csv* in this applications. In particular, the dependent variable is *MedVisPrevOr*, which is an ordered variable equal to 1 if the individual did not visit a physician for preventive reasons, 2 if the individual visited once in that year, and so on, until it is equal to 6 for visiting five or more times. The latter category is 1.6% of the sample. Observe that the dependent variable has six categories.

In this example, the set of regressors is given by *SHI*, which an indicator of being in the subsidized health care system (1 means being in the system), sex (*Female*), age (linear and squared), socioeconomic conditions indicator (*Est2* and *Est3*), the lowest is tha baseline category, self perception of health status (*Fair*, *Good* and *Excellent*), where *Bad* is the baseline, and education level, primary (*PriEd*), high school (*HighEd*), vocational (*VocEd*), and university (*UnivEd*), *no education* is the baseline category.

We ran this application with 50,000 MCMC iterations plus 10,000 as burn-in, and thinning parameter equal to 5. This setting means 10,000 effective posterior draws. We set $\boldsymbol{\beta}_0 = \mathbf{0}_{11}$, $\boldsymbol{B}_0 = 1000\boldsymbol{I}_{11}$, $\boldsymbol{\gamma}_0 = \mathbf{0}_4$, $\boldsymbol{\Gamma}_0 = \boldsymbol{I}_4$, and the tuning parameter is 1.

We can run the ordered probit models in our GUI following the steps in the Algorithm A5.

The following **R** code shows how to perform inference in this model using the command *rordprobitGibbs* from the *bayesm* library, which is the command that our GUI uses.

Algorithm A5 Ordered probit models

1: Select *Univariate Models* on the top panel
2: Select *Ordered Probit* model using the left radio button
3: Upload the dataset selecting first if there is header in the file, and the kind of separator in the *csv* file of the dataset (comma, semicolon, or tab). Then, use the *Browse* button under the **Choose File** legend. You should see a preview of the dataset
4: Select MCMC iterations, burn-in and thinning parameters using the *Range sliders*
5: Select dependent and independent variables using the *Formula builder* table
6: Click the *Build formula* button to generate the formula in **R** syntax. Remember that this formula must have -1 to omit the intercept in the specification.
7: Set the hyperparameters: mean vectors and covariance matrices. This step is not necessary as by default our GUI uses non-informative priors
8: Select the tuning parameter for the Metropolis-Hastings algorithm
9: Click the *Go!* button
10: Analyze results
11: Download posterior chains and diagnostic plots using the *Download Posterior Chains* and *Download Posterior Graphs* buttons

R. code. Determinants of preventive health care visits

```
rm(list = ls())
set.seed(010101)
Data <- read.csv("DataApplications/2HealthMed.csv", sep = ",
    ", header = TRUE, fileEncoding = "latin1")
attach(Data)
y <- MedVisPrevOr
# MedVisPrevOr: Oredered variable for preventive visits to
    doctors in one year: 1 (none), 2 (once), ... 6 (five or
    more)
X <- cbind(SHI, Female, Age, Age2, Est2, Est3, Fair, Good,
    Excellent, PriEd, HighEd, VocEd, UnivEd)
k <- dim(X)[2]
L <- length(table(y))
# Hyperparameters
b0 <- rep(0, k); c0 <- 1000; B0 <- c0*diag(k)
gamma0 <- rep(0, L-2); Gamma0 <- diag(L-2)
# MCMC parameters
mcmc <- 60000+1; thin <- 5; tuningPar <- 1/(L-2)^0.5
DataApp <- list(y = y, X = X, k = L)
Prior <- list(betabar = b0, A = solve(B0), dstarbar = gamma0
    , Ad = Gamma0)
mcmcpar <- list(R = mcmc, keep = 5, s = tuningPar)
PostBeta <- bayesm::rordprobitGibbs(Data = DataApp, Prior =
    Prior, Mcmc = mcmcpar)
```

R. code. Determinants of preventive health care visits, results

```
BetasPost <- coda::mcmc(PostBeta[["betadraw"]])
colnames(BetasPost) <- c("SHI", "Female", "Age", "Age2", "
    Est2", "Est3", "Fair", "Good", "Excellent", "PriEd", "
    HighEd", "VocEd", "UnivEd")
summary(BetasPost)
Iterations = 1:12000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 12000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
Mean         SD  Naive SE Time-series SE
SHI         0.0654824 2.281e-02 2.082e-04      3.357e-04
Female     -0.0374788 1.908e-02 1.742e-04      1.742e-04
Age         0.0190336 1.869e-03 1.706e-05      4.576e-05
Age2       -0.0002328 2.438e-05 2.225e-07      6.690e-07
Est2        0.0949445 2.226e-02 2.032e-04      4.659e-04
Est3       -0.1383965 3.411e-02 3.114e-04      3.459e-04
Fair        0.6451828 5.375e-02 4.907e-04      3.924e-03
Good        0.7343932 4.955e-02 4.523e-04      4.491e-03
Excellent   0.9826531 6.393e-02 5.836e-04      5.261e-03
PriEd       0.0309418 2.376e-02 2.169e-04      2.221e-04
HighEd     -0.1805753 2.910e-02 2.656e-04      3.456e-04
VocEd       0.1395760 9.640e-02 8.800e-04      9.291e-04
UnivEd     -0.2218120 1.189e-01 1.086e-03      1.086e-03
2. Quantiles for each variable:
          2.5%        25%         50%         75%       97.5%
SHI         0.02090   0.04995   0.06540   0.08085   0.11021
Female     -0.07463  -0.05042  -0.03777  -0.02456   0.00023
Age         0.01550   0.01781   0.01902   0.02023   0.02268
Age2       -0.00028  -0.00024  -0.00023  -0.00021  -0.00018
Est2        0.05149   0.08004   0.09482   0.10968   0.13933
Est3       -0.20559  -0.16144  -0.13815  -0.11563  -0.07179
Fair        0.55799   0.61295   0.64148   0.67268   0.74395
Good        0.66690   0.70808   0.73032   0.75406   0.81064
Excellent   0.88919   0.94770   0.97836   1.01026   1.08460
PriEd      -0.01584   0.01493   0.03101   0.04718   0.07732
HighEd     -0.23782  -0.20035  -0.18021  -0.16073  -0.12435
VocEd      -0.04911   0.07474   0.13811   0.20414   0.33331
UnivEd     -0.45381  -0.30239  -0.22193  -0.14148   0.00863
# Convergence diagnostics
coda::geweke.diag(BetasPost)
coda::raftery.diag(BetasPost,q=0.5,r=0.05,s = 0.95)
coda::heidel.diag(BetasPost)
# Cut offs
Cutoffs <- PostBeta[["cutdraw"]]
summary(Cutoffs)
coda::geweke.diag(Cutoffs)
coda::heidel.diag(Cutoffs)
coda::raftery.diag(Cutoffs[,-1],q=0.5,r=0.05,s = 0.95)
```

The results suggest that older individuals (at decreasing rate) in the subsidized health program, characterized in the second socioeconomic status with increasing good self perception of health condition, and not having high school as their highest education degree, have a higher probability of visiting a physician for preventive health aims. Convergence diagnostics look well, except for the self health perception draws.

We also got the posterior estimates of the cutoffs in the ordered probit model. These estimates are necessary to calculate the probability that an individual is in a specific category of visiting physicians. Due to identification restrictions, the first cutoff is set equal to 0. That is why we have *NaN* values in [36] and [42] tests, and we observe only four values in the [72] test, which correspond to the remaining free cutoffs. It seems that these cutoff estimates have some convergence issues when taking as diagnostic tool the [72] test. Their dependence factors are also very high.

7.7 Negative binomial model

The dependent variable in the negative binomial model is a nonnegative integer or count. In contrast to the Poisson model, the negative binomial model takes into account over-dispersion. The Poisson model has equal mean and variance (equi-dispersion).

We assume that $y_i \overset{i.n.d.}{\sim} NB(\gamma, \theta_i)$, that is, the density function for individual i is $\frac{\Gamma(y_i+\gamma)}{\Gamma(\gamma)y_i!}(1-\theta_i)^{y_i}\theta_i^{\gamma}$, where the success probability is $\theta_i = \frac{\gamma}{\lambda_i+\gamma}$, $\lambda_i = \exp\left\{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right\}$ is the mean, and $\gamma = \exp\{\alpha\}$ is the target for number of successful trials, or dispersion parameter.

We assume independent priors for this model are $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$ and $\alpha \sim G(\alpha_0, \delta_0)$.[13]

This model does not have standard conditional posterior distributions, so [77] use a random-walk Metropolis–Hastings algorithm where the proposal distribution for $\boldsymbol{\beta}$ is Gaussian centered at the current stage with covariance matrix $s_{\boldsymbol{\beta}}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ where $s_{\boldsymbol{\beta}}$ is a tuning parameter and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ is the maximum likelihood covariance estimator. In addition, the proposal for α is normal centered at the current value, with variance $s_\alpha^2\hat{\sigma}_\alpha^2$ where s_α is a tuning parameter and $\hat{\sigma}_\alpha^2$ is the maximum likelihood variance estimator.

Example: Simulation exercise

Let's do a simulation exercise to check the performance of the M-H algorithms in the negative binomial model. There are two regressors, $x_{i1} \sim U(0,1)$ and $x_{i1} \sim N(0,1)$, and the intercept. The dispersion parameter is $\gamma = \exp\{1.2\})$, and $\boldsymbol{\beta} = [1\ 1\ 1]$. The sample size is 1,000.

[13] G denotes a gamma density.

We run this simulation using 10,000 MCMC iterations, a burn-in equal to 1,000, and a thinning parameter equal to 5. We set vague priors for the location parameters, particularly, $\boldsymbol{\beta}_0 = \mathbf{0}_3$ and $\boldsymbol{B}_0 = 1000\boldsymbol{I}_3$, and $\alpha_0 = 0.5$ and $\delta_0 = 0.1$, which are the default values in the *rnegbinRw* command from *bayesm* package in **R**. In addition, the tuning parameters of the Metropolis–Hastings algorithms are $s_\beta = 2.93/k^{1/2}$ and $s_\alpha = 2.93$, which are also the default parameters in *rnegbinRw*, k is the number of location parameters.

We can run the negative binomial models in our GUI following the steps in the Algorithm A6.

Algorithm A6 Negative binomial models

1: Select *Univariate Models* on the top panel
2: Select *Negative Binomial (Poisson)* model using the left radio button
3: Upload the dataset selecting first if there is header in the file, and the kind of separator in the *csv* file of the dataset (comma, semicolon, or tab). Then, use the *Browse* button under the **Choose File** legend. You should see a preview of the dataset
4: Select MCMC iterations, burn-in and thinning parameters using the *Range sliders*
5: Select dependent and independent variables using the *Formula builder* table
6: Click the *Build formula* button to generate the formula in **R** syntax.
7: Set the hyperparameters: mean vector, covariance matrix, shape and scale parameters. This step is not necessary as by default our GUI uses non-informative priors
8: Select the tuning parameters for the Metropolis-Hastings algorithms
9: Click the *Go!* button
10: Analyze results
11: Download posterior chains and diagnostic plots using the *Download Posterior Chains* and *Download Posterior Graphs* buttons

The following **R** code shows how to perform inference in the negative binomial model programming the M-H algorithms from scratch. We ask to estimate this example using the *rnegbinRw* command in exercise 8.

We observe from the results that all 95% credible intervals encompass the population parameters, and the posterior means are very close to the population parameters.

R. code. Simulation of the negative binomial model

```
rm(list = ls())
set.seed(010101)
N <- 2000 # Sample size
x1 <- runif(N); x2 <- rnorm(N)
X <- cbind(1, x1, x2); k <- dim(X)[2]; B <- rep(1, k)
alpha <- 1.2; gamma <- exp(alpha); lambda <- exp(X%*%B)
y <- rnbinom(N, mu = lambda, size = gamma)
# log likelihood
logLik <- function(par){
  alpha <- par[1]; beta <- par[2:(k+1)]
  gamma <- exp(alpha)
  lambda <- exp(X%*%beta)
  logLikNB <- sum(sapply(1:N, function(i){dnbinom(y[i], size
     = gamma, mu = lambda[i], log = TRUE)}))
  return(-logLikNB)
}
# Parameters: Proposal
par0 <- rep(0.5, k+1)
res.optim <- optim(par0, logLik, method="BFGS", hessian=TRUE
    )
res.optim$par
res.optim$convergence
Covar <- solve(res.optim$hessian)
CovarBetas <- Covar[2:(k+1),2:(k+1)]
VarAlpha <- Covar[1:1]
# Hyperparameters: Priors
B0 <- 1000*diag(k); b0 <- rep(0, k)
alpha0 <- 0.5; delta0 <- 0.1
```

R. code. Simulation of the negative binomial model, M-H algorithm

```
# Metropolis-Hastings function
MHfunction <- function(iter, sbeta, salpha){
  Beta <- rep(0, k);  Acept1 <- NULL; Acept2 <- NULL
  BetasPost <- matrix(NA, iter, k); alpha <- 1
  alphaPost <- rep(NA, iter); par <- c(alpha, Beta)
  pb <- winProgressBar(title = "progress bar", min = 0, max
    = iter, width = 300)
  for(s in 1:iter){
    LogPostBeta <- -logLik(par) + dgamma(alpha, shape =
    alpha0, scale = delta0, log = TRUE) + mvtnorm::dmvnorm(
    Beta, mean = b0, sigma = B0, log = TRUE)
    BetaC <- c(MASS::mvrnorm(1, mu = Beta, Sigma = sbeta^2*
    CovarBetas))
    parC <- c(alpha, BetaC)
    LogPostBetaC <- -logLik(parC) + dgamma(alpha, shape =
    alpha0, scale = delta0, log = TRUE) +  mvtnorm::dmvnorm(
    BetaC, mean = b0, sigma = B0, log = TRUE)
    alpha1 <- min(exp((LogPostBetaC - mvtnorm::dmvnorm(BetaC
    , mean = Beta, sigma = sbeta^2*CovarBetas, log = TRUE))
    -(LogPostBeta - mvtnorm::dmvnorm(Beta, mean = Beta,
    sigma = sbeta^2*CovarBetas, log = TRUE))),1)
    u1 <- runif(1)
    if(u1 <= alpha1){Acept1i <- 1; Beta <- BetaC}else{
      Acept1i <- 0; Beta <- Beta
    }
    par <- c(alpha, Beta)
    LogPostBeta <- -logLik(par) + dgamma(alpha, shape =
    alpha0, scale = delta0, log = TRUE) + mvtnorm::dmvnorm(
    Beta, mean = b0, sigma = B0, log = TRUE)
    alphaC <- rnorm(1, mean = alpha, sd = salpha*VarAlpha
    ^0.5)
    parC <- c(alphaC, Beta)
    LogPostBetaC <- -logLik(parC) + dgamma(alphaC, shape =
    alpha0, scale = delta0, log = TRUE) +  mvtnorm::dmvnorm(
    Beta, mean = b0, sigma = B0, log = TRUE)
    alpha2 <- min(exp((LogPostBetaC - dnorm(alphaC, mean =
    alpha, sd = salpha*VarAlpha^0.5, log = TRUE))-(
    LogPostBeta - dnorm(alpha, mean = alpha, sd = salpha*
    VarAlpha^0.5, log = TRUE))),1)
    u2 <- runif(1)
    if(u2 <= alpha2){Acept2i <- 1; alpha <- alphaC}else{
      Acept2i <- 0; alpha <- alpha
    }

    BetasPost[s, ] <- Beta; alphaPost[s] <- alpha
    Acept1 <- c(Acept1, Acept1i); Acept2 <- c(Acept2,
    Acept2i)
    setWinProgressBar(pb, s, title=paste( round(s/iter*100,
    0),"% done"))
  }
  close(pb)
  AcepRateBeta <- mean(Acept1); AcepRateAlpha <- mean(Acept2
    )
  Results <- list(AcepRateBeta = AcepRateBeta, AcepRateAlpha
     = AcepRateAlpha, BetasPost = BetasPost, alphaPost =
    alphaPost)
  return(Results)
}
```

R. code. Simulation of the negative binomial model, results

```
# MCMC parameters
mcmc <- 10000
burnin <- 1000
thin <- 5
iter <- mcmc + burnin
keep <- seq(burnin, iter, thin)
sbeta <- 2.93/sqrt(k); salpha <- 2.93
# Run M-H
ResultsPost <- MHfunction(iter = iter, sbeta = sbeta, salpha
     = salpha)
ResultsPost$AcepRateBeta
ResultsPost$AcepRateAlpha
summary(coda::mcmc(ResultsPost$BetasPost[keep[-1], ]))
Iterations = 1:2000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 2000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
      Mean      SD  Naive SE Time-series SE
[1,] 1.0270 0.04799 0.0010730      0.0014727
[2,] 0.9981 0.07752 0.0017333      0.0024262
[3,] 0.9677 0.02343 0.0005239      0.0007182
2. Quantiles for each variable:
      2.5%    25%    50%    75% 97.5%
var1 0.9343 0.9943 1.0255 1.0592 1.122
var2 0.8445 0.9448 0.9980 1.0520 1.144
var3 0.9242 0.9512 0.9678 0.9839 1.013
summary(coda::mcmc(ResultsPost$alphaPost[keep[-1]]))
Iterations = 1:2000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 2000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
      Mean             SD       Naive SE Time-series SE
1.282664       0.058769       0.001314       0.001427
2. Quantiles for each variable:
2.5%   25%   50%   75% 97.5%
1.173 1.242 1.282 1.320 1.407
```

7.8 Tobit model

The dependent variable is partially observed in Tobit models due to sampling schemes, whereas the regressors are completely observed. In particular,

$$y_i = \left\{ \begin{array}{c} L\ , \quad y_i^* < L \\ y_i^*\ ,\ L \leq y_i^* < U \\ U\ , \quad y_i^* \geq U \end{array} \right\},$$

where $y_i^* \overset{i.n.d.}{\sim} N(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$.[14]

We use conjugate independent priors $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$ and $\sigma^2 \sim IG(\alpha_0/2, \delta_0/2)$, and data augmentation using $\boldsymbol{y}_C^*$ such that $y_{C_i}^* \overset{i.n.d.}{\sim} N(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, $y_{C_i} = \left\{ y_{C_i^L}^* \cup y_{C_i^U}^* \right\}$ are lower and upper censored data. This allows implementing the Gibbs sampling algorithm [18]. Then,

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{y^*} | \boldsymbol{y}, \boldsymbol{X}) \propto \prod_{i=1}^{N} & \left[\mathbb{1}_{y_i = L} \mathbb{1}_{y_{C_i^L}^* < L} + \mathbb{1}_{L \leq y_i < U} + \mathbb{1}_{y_i = U} \mathbb{1}_{y_{C_i^U}^* \geq U} \right] \\ & \times N(y_i^* | \boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2) \times N(\boldsymbol{\beta} | \boldsymbol{\beta}_0, \boldsymbol{B}_0) \times IG(\sigma^2 | \alpha_0/2, \delta_0/2) \end{aligned}$$

The posterior distributions are

$$y_{C_i}^* | \boldsymbol{\beta}, \sigma^2, \boldsymbol{y}, \boldsymbol{X} \sim \left\{ \begin{array}{l} TN_{(-\infty, L)}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)\ ,\ y_i = L \\ TN_{[U, \infty)}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2) \quad ,\ y_i = U \end{array} \right\},$$

$$\boldsymbol{\beta} | \sigma^2, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\beta}_n, \sigma^2 \boldsymbol{B}_n),$$

$$\sigma^2 | \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X} \sim IG(\alpha_n/2, \delta_n/2),$$

where $\boldsymbol{B}_n = (\boldsymbol{B}_0^{-1} + \sigma^{-2} \boldsymbol{X}^\top \boldsymbol{X})^{-1}$, $\boldsymbol{\beta}_n = \boldsymbol{B}_n (\boldsymbol{B}_0^{-1} \boldsymbol{\beta}_0 + \sigma^{-2} \boldsymbol{X}^\top \boldsymbol{y})$, $\alpha_n = \alpha_0 + N$ and $\delta_n = \delta_0 + (\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\beta})$.

7.9 Summary

We present ...

[14] We can set L or U equal to $-\infty$ or ∞ to model data censored in just one side.

7.10 Exercises

1. Get the posterior conditional distributions of the Gaussian linear model assuming independent priors $\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta}) \times \pi(\sigma^2)$, where $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{B}_0)$ and $\sigma^2 \sim IG(\alpha_0/2, \delta_0/2)$.
2. Show that the posterior conditional distributions of the Gaussian linear model with heteroskedasticity assuming independent priors $\pi(\boldsymbol{\beta}, \sigma^2, \tau) = \pi(\boldsymbol{\beta}) \times \pi(\sigma^2) \times \prod_{i=1}^{N} \pi(\tau_i)$, where $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{B}_0)$, $\sigma^2 \sim IG(\alpha_0/2, \delta_0/2)$ and $\tau_i \sim G(v/2, v/2)$ are $\boldsymbol{\beta}|\sigma^2, \tau, \mathbf{y}, \mathbf{X} \sim N(\boldsymbol{\beta}_n, \mathbf{B}_n)$, $\sigma^2|\boldsymbol{\beta}, \tau, \mathbf{y}, \mathbf{X} \sim IG(\alpha_n, \delta_n)$ and $\tau_i|\boldsymbol{\beta}, \sigma^2, \mathbf{y}, \mathbf{X} \sim G(v_{1n}, v_{2in})$, where $\tau = [\tau_1, \ldots, \tau_n]^\top$, $\mathbf{B}_n = (\mathbf{B}_0^{-1} + \sigma^{-2}\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1}$, $\boldsymbol{\beta}_n = \mathbf{B}_n(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{y})$, $\alpha_n = \alpha_0 + N$, $\delta_n = \delta_0 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Psi} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, $v_{1n} = v + 1$, $v_{2in} = v + \sigma^{-2}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$, and $\boldsymbol{\Psi} = \text{diagonal}\,\{\tau_i\}$.
3. **The market value of soccer players in Europe continues**

   Use the setting of the previous exercise to perform inference using a Gibbs sampling algorithm of the the market value of soccer players in Europe setting $v = 5$ and same other hyperparameters as the homoscedastic case. Is there any meaningful difference for the coefficient associated with the national team compared to the application in the homoscedastic case?
4. **Example: Determinants of hospitalization continues**

   Program a Gibbs sampling algorithm in the application of determinants of hospitalization.
5. **Choice of the fishing mode continues**

   Run the Algorithm A3 of the book to show the results of the Geweke [36], Raftery [72] and Heidelberger [42] tests using our GUI.
6. **Simulation exercise of the multinomial logit model continues**

   Perform inference in the simulation of the multinomial logit model using the command *rmnlIndepMetrop* from the *bayesm* package of **R** and using our GUI.
7. **Simulation of the ordered probit model**

   Simulate an ordered probit model where the first regressor distributes $N(6, 5)$ and the second distributes $G(1, 1)$, the location parameter is $\boldsymbol{\beta} = [0.5 \ -0.25 \ 0.5]$, and the cutoffs is the vector $\boldsymbol{\alpha} = [0 \ 1 \ 2.5]$. Program from scratch a Metropolis-within-Gibbs sampling algorithm for this simulation.
8. **Simulation of the negative binomial model continues**

   Perform inference in the simulation of the negative binomial model using the *bayesm* package in **R** software.

8

Multivariate models

9

Time series models

10

Panel data models

11

Bayesian model average

Part III

Advanced methods: Theory, applications and programming

12

Hierarchical models

12.1 Finite mixtures

12.2 Direchlet processes

13

Causal inference

14

Machine learning

15

Further topics

Bibliography

[1] J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[2] R. Baath. *Package bayesboot*, 2018.

[3] M. Bayarri and J. Berger. P–values for composite null models. *Journal of American Statistical Association*, 95:1127–1142, 2000.

[4] M. J. Bayarri and J. Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

[5] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–416, 1763.

[6] Thomas Bayes. LII. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, 53:370–418, 1763.

[7] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature human behaviour*, 2(1):6–10, 2018.

[8] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, third edition edition, 1993.

[9] J. Berger. The case for objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

[10] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[11] J. Bernardo and A. Smith. *Bayesian Theory*. Wiley, Chichester, 1994.

[12] Peter J Bickel and Joseph A Yahav. Some contributions to the asymptotic theory of bayes solutions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11(4):257–276, 1969.

[13] G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71:791–799, 1976.

[14] George EP Box. Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier, 1979.

[15] Colin Cameron and Pravin Trivedi. *Microeconometrics: Methods and Applications*. Cambridge, 2005.

[16] W. Chang. *Web Application Framework for R: Package shiny*. R Studio, 2018.

[17] V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115:293–346, 2003.

[18] S. Chib. Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, 51:79–99, 1992.

[19] Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432):1313–1321, 1995.

[20] Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

[21] M. Clyde and E. George. Model uncertatinty. *Statistical Science*, 19(1):81–94, 2004.

[22] A. P. Dawid, M. Musio, and S. E. Fienberg. From statistical evidence to evidence of causality. *Bayesian Analysis*, 11(3):725–752, 2016.

[23] de Finetti. Foresight: its logical laws, its subjective sources. In H. E. Kyburg and H. E. Smokler, editors, *Studies in Subjective Probability*. Krieger, New York, 1937. p.55–118.

[24] M. H. DeGroot. *Probability and statistics*. Addison-Wesley Publishing Co., London, 1975.

[25] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281, 1979.

[26] Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

[27] R. Fisher. *Statistical Methods for Research Workers*. Hafner, New York, 13th edition, 1958.

[28] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[29] Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

[30] A. Gelman and X. Meng. Model checking and model improvement. In Gilks, Richardson, and Speigelhalter, editors, *In Markov chain Monte Carlo in practice*. Springer US, 1996. Chapter 6, pp. 157–196.

[31] A. Gelman, X. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

[32] Andrew Gelman, John B Carlin, Hal S Stern, David Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2021.

[33] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[34] Andrew Gelman and Guido Imbens. Why ask why? forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research, 2013.

[35] S Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[36] J. Geweke. *Bayesian Statistics*, chapter Evaluating the accuracy of sampling-based approaches to calculating posterior moments. Clarendon Press, Oxford, UK., 1992.

[37] John Geweke. *Contemporary Bayesian econometrics and statistics*, volume 537. John Wiley & Sons, 2005.

[38] I. J. Good. The bayes/non bayes compromise: A brief review. *Journal of the American Statistical Association*, 87(419):597–606, September 1992.

[39] S. N. Goodman. Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of internal medicine*, 130(12):995–1004, 1999.

[40] Edward Greenberg. *Introduction to Bayesian econometrics*. Cambridge University Press, 2012.

[41] W. Hastings. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57:97–109, 1970.

[42] P. Heidelberger and P. D. Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144, 1983.

[43] H. Jeffreys. Some test of significance, treated by the theory of probability. *Proccedings of the Cambridge philosophy society*, 31:203–222, 1935.

[44] H. Jeffreys. *Theory of Probability*. Oxford University Press, London, 1961.

[45] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[46] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[47] G. Karabatsos. A menu-driven software package of Bayesian nonparametric (and parametric) mixed models for regression analysis and density estimation. *Behavior Research Methods*, 49:335–362, 2016.

[48] R. Kass. Statistical inference: the big picture. *Statistical science*, 26(1):1–9, 2011.

[49] Robert E. Kass and Adrian E. Raftery. Bayes factorss. *Journal of American Statistical Association*, 90(430):773–795, 1995.

[50] Gary M Koop. *Bayesian econometrics*. John Wiley & Sons Inc., 2003.

[51] Tony Lancaster. *An introduction to modern Bayesian econometrics*. Blackwell Oxford, 2004.

[52] P. Laplace. *Théorie Analytique des Probabilités*. Courcier, 1812.

[53] Pierre Simon Laplace. Mémoire sur la probabilité de causes par les évenements. *Mémoire de l'académie royale des sciences*, 1774.

[54] E.L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, second edition edition, 2003.

[55] Alex Lenkoski, Anna Karl, and Andreas Neudecker. *Package ivbma*, 2013.

[56] D. V. Lindley. The philosophy of statistics. *The Statistician*, 49(3):293–337, 2000.

[57] D. V. Lindley and L. D. Phillips. Inference for a Bernoulli process (a Bayesian view). *American Statistician*, 30:112–119, 1976.

[58] Dennis V Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

[59] Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park. MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software*, 42(9):1–21, 2011.

[60] Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park. *Package MCMCpack*, 2018.

[61] R. McCulloch and P. Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64:207–240, 1994.

[62] Sharon Bertsch McGrayne. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of C.* Yale University Press, 2011.

[63] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys*, 21:1087–1092, 1953.

[64] J. Neyman and E. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A*, 231:289–337, 1933.

[65] G. Parmigiani and L. Inoue. *Decision theory principles and approaches.* John Wiley & Sons, 2008.

[66] Luis Pericchi and Carlos Pereira. Adaptative significance levels using optimal decision rules: Balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics*, 2015.

[67] Giovanni Petris, Sonia Petrone, and Patrizia Campagnoli. Dynamic linear models. In *Dynamic Linear Models with R*, pages 31–84. Springer, 2009.

[68] Martyn Plummer, Nicky Best, Kate Cowles, Karen Vines, Deepayan Sarkar, Douglas Bates, Russell Almond, and Arni Magnusson. *Output Analysis and Diagnostics for MCMC*, 2016.

[69] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021.

[70] A. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

[71] Adrian Raftery, Jennifer Hoeting, Chris Volinsky, Ian Painter, and Ka Yee Yeung. *Package BMA*, 2012.

[72] A.E. Raftery and S.M. Lewis. One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7:493–497, 1992.

[73] A. Ramírez Hassan, J. Cardona Jiménez, and R. Cadavid Montoya. The impact of subsidized health insurance on the poor in Colombia: Evaluating the case of Medellín. *Economia Aplicada*, 17(4):543–556, 2013.

[74] F. Ramsey. Truth and probability. In Routledge and Kegan Paul, editors, *The Foundations of Mathematics and other Logical Essays.* New York: Harcourt, Brace and Company, London, 1926. Ch. VII, p.156–198.

[75] A. Ramírez-Hassan and M. Graciano-Londoño. A guided tour of Bayesian regression. Technical report, Universidad EAFIT, 2020.

[76] P. Rossi. *Package bayesm*, 2017.

[77] Peter E Rossi, Greg M Allenby, and Rob McCulloch. *Bayesian statistics and marketing*. John Wiley & Sons, 2012.

[78] L. J. Savage. *The foundations of statistics*. John Wiley & Sons, Inc., New York, 1954.

[79] Robert Schlaifer and Howard Raiffa. *Applied statistical decision theory*. Wiley New York, 1961.

[80] Thomas Sellke, MJ Bayarri, and James O Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

[81] Steve Selvin. A problem in probability (letter to the editor). *The American Statistician*, 11(1):67–71, 1975.

[82] Steve Selvin. A problem in probability (letter to the editor). *The American Statistician*, 11(3):131–134, 1975.

[83] M. Serna Rodríguez, A. Ramírez Hassan, and A. Coad. Uncovering value drivers of high performance soccer players. *Journal of Sport Economics*, 20(6):819–849, 2019.

[84] A. F. M. Smith. A General Bayesian Linear Model. *Journal of the Royal Statistical Society. Series B (Methodological).*, 35(1):67–75, 1973.

[85] Stan Development Team. shinystan: Interactive visual and numerical diagnostics and posterior analysis for Bayesian models., 2017. R package version 2.3.0.

[86] Stephen Stigler. Richard price, the first bayesian. *Statistical Science*, 33(1):117–125, 2018.

[87] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

[88] Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728, 1994.

[89] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[90] Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p–values: context, process and purpose. *The American Statistician*, 2016.

[91] P. Woodward. BugsXLA: Bayes for the common man. *Journal of Statistical Software*, 14(5):1–18, 2005.

[92] A. Zellner. *Introduction to Bayesian inference in econometrics*. John Wiley & Sons Inc., 1996.

[93] S. Ziliak. Guinnessometrics; the Economic Foundation of student's t. *Journal of Economic Perspectives*, 22(4):199–216, 2008.

TABLE 15.1
Libraries and commands in BEsmarter GUI.

Univariate models			
Model	**Library**	**Command**	**Reference**
Normal	MCMCpack	MCMCregress	[60]
Logit	MCMCpack	MCMClogit	[60]
Probit	bayesm	rbprobitGibbs	[76]
Multinomial(Mixed) Probit	bayesm	rmnpGibbs	[76]
Multinomial(Mixed) Logit	bayesm	rmnlIndepMetrop	[76]
Ordered Probit	bayesm	rordprobitGibbs	[76]
Negative Binomial(Poisson)	bayesm	rnegbinRw	[76]
Tobit	MCMCpack	MCMCtobit	[60]
Quantile	MCMCpack	MCMCquantreg	[60]
Multivariate models			
Model	**Library**	**Command**	**Reference**
Multivariate	bayesm	rmultireg	[76]
Seemingly Unrelated Regression	bayesm	rsurGibbs	[76]
Instrumental Variable	bayesm	rivGibbs	[76]
Bivariate Probit	bayesm	rmvpGibbs	[76]
Hierarchical longitudinal models			
Model	**Library**	**Command**	**Reference**
Normal	MCMCpack	MCMChregress	[60]
Logit	MCMCpack	MCMChlogit	[60]
Poisson	MCMCpack	MCMChpoisson	[60]
Bayesian Bootstrap			
Model	**Library**	**Command**	**Reference**
Bayesian bootstrap	bayesboot	bayesboot	[2]
Bayesian model averaging			
Model	**Library**	**Command**	**Reference**
Normal (BIC)	BMA	bic.glm	[71]
Normal (MC^3)	BMA	MC3.REG	[71]
Normal (instrumental variables)	ivbma	ivbma	[55]
Logit (BIC)	BMA	bic.glm	[71]
Gamma (BIC)	BMA	bic.glm	[71]
Poisson (BIC)	BMA	bic.glm	[71]
Diagnostics			
Diagnostic	**Library**	**Command**	**Reference**
Trace plot	coda	traceplot	[68]
Autocorrelation plot	coda	autocorr.plot	[68]
Geweke test	coda	geweke.diag	[68]
Raftery & Lewis test	coda	raftery.diag	[68]
Heidelberger & Welch test	coda	heidel.diag	[68]

TABLE 15.2
Datasets templates in folder *DataSim*.

Univariate models		
Model	**Data set file**	**Data set simulation**
Normal	11SimNormalmodel.csv	11SimNormal.R
Logit	12SimLogitmodel.csv	12SimLogit
Probit	13SimProbitmodel.csv	13SimProbit.R
Multinomial(Mixed) Probit	14SimMultProbmodel.csv	14SimMultinomialProbit.R
Multinomial(Mixed) Logit	15SimMultLogitmodel.csv	15SimMultinomialLogit.R
Ordered Probit	16SimOrderedProbitmodel.csv	16SimOrderedProbit.R
Negative Binomial(Poisson)	17SimNegBinmodel.csv	17SimNegBin.R
Tobit	18SimTobitmodel.csv	18SimTobit.R
Quantile	19SimQuantilemodel.csv	19SimQuantile.R
Multivariate models		
Model	**Data set file**	**Data set simulation**
Multivariate	21SimMultivariate.csv	21SimMultReg.R
Seemingly Unrelated Regression	22SimSUR.csv	22SimSUR.R
Instrumental Variable	23SimIV.csv	23SimIV.R
Bivariate Probit	24SimMultProbit.csv	24SimMultProbit.R
Hierarchical longitudinal models		
Model	**Data set file**	**Data set simulation**
Normal	31SimLogitudinalNormal.csv	31SimLogitudinalNormal.R
Logit	32SimLogitudinalLogit.csv	32SimLogitudinalLogit.R
Poisson	33SimLogitudinalPoisson.csv	33SimLogitudinalPoisson.R
Bayesian Bootstrap		
Model	**Data set file**	**Data set simulation**
Bayesian bootstrap	41SimBootstrapmodel.csv	41SimBootstrapmodel.R
Bayesian model averaging		
Model	**Data set file**	**Data set simulation**
Normal (BIC)	511SimNormalBMA.csv	511SimNormalBMA.R
Normal (MC^3)	512SimNormalBMA.csv	512SimNormalBMA.R
Normal (instrumental variables)	513SimNormalBMAivYXW.csv 513SimNormalBMAivZ.csv	513SimNormalBMAiv.R
Logit (BIC)	52SimLogitBMA.csv	52SimLogitBMA.R
Gamma (BIC)	53SimGammaBMA.csv	53SimGammaBMA.R
Poisson (BIC)	53SimPoissonBMA.csv	53SimPoissonBMA.R

TABLE 15.3
Real datasets in folder *DataApp*.

Univariate models		
Model	**Data set file**	**Dependent variable**
Normal	1ValueFootballPlayers.csv	log(Value)
Logit	2HealthMed.csv	Hosp
Probit	2HealthMed.csv	Hosp
Multinomial(Mixed) Probit	Fishing.csv	mode
Multinomial(Mixed) Logit	Fishing.csv	mode
Ordered Probit	2HealthMed.csv	MedVisPrevOr
Negative Binomial(Poisson)	2HealthMed.csv	MedVisPrev
Tobit	1ValueFootballPlayers.csv	log(ValueCens)
Quantile	1ValueFootballPlayers.csv	log(Value)
Multivariate models		
Model	**Data set file**	**Dependent variable**
Multivariate	4Institutions.csv	logpcGDP95 and PAER
Seemingly Unrelated Regression	5Institutions.csv	logpcGDP95 and PAER
Instrumental Variable	6Institutions.csv	logpcGDP95 and PAER
Bivariate Probit	7HealthMed.csv	$y = [\text{Hosp SHI}]'$
Hierarchical longitudinal models		
Model	**Data set file**	**Dependent variable**
Normal	8PublicCap.csv	log(gsp)
Logit	9VisitDoc.csv	DocVis
Poisson	9VisitDoc.csv	DocNum
Bayesian Bootstrap		
Model	**Data set file**	**Dependent variable**
Bayesian bootstrap	1ValueFootballPlayers.csv	log(Value)
Bayesian model averaging		
Model	**Data set file**	**Dependent variable**
Normal (BIC)	10ExportDiversificationHHI.csv	avghhi
Normal (MC^3)	10ExportDiversificationHHI.csv	avghhi
Normal (instrumental variables)	11ExportDiversificationHHI.csv 12ExportDiversificationHHIInstr.csv	avghhi and avglgdpcap
Logit (BIC)	13InternetMed.csv	internet
Gamma (BIC)	14ValueFootballPlayers.csv	log market value
Poisson (BIC)	15Fertile2.csv	ceb