Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Springer Texts in Statistics

Athreya/Lahiri: Measure Theory and Probability Theory
Bilodeau/Brenner: Theory of Multivariate Statistics
Brockwell/Davis: An Introduction to Time Series and Forecasting
Carmona: Statistical Analysis of Financial Data in S-PLUS
Chow/Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization, Second Edition
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Davis: Statistical Methods for the Analysis of Repeated Measurements
Dean/Voss: Design and Analysis of Experiments
Dekking/Kraaikamp/Lopuhaä/Meester: A Modern Introduction to Probability and Statistics
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modeling, Second Edition
Everitt: An R and S-PLUS Companion to Multivariate Analysis
Ghosh/Delampady/Samanta: An Introduction to Bayesian Analysis
Gut: Probability: A Graduate Course
Heiberger/Holland: Statistical Analysis and Data Display; An Intermediate Course with Examples in S-PLUS, R, and SAS
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Karr: Probability
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lange: Applied Probability
Lange: Optimization
Lehmann: Elements of Large Sample Theory
Lehmann/Romano: Testing Statistical Hypotheses, Third Edition
Lehmann/Casella: Theory of Point Estimation, Second Edition
Marin/Robert: Bayesian Core: A Practical Approach to Computational Bayesian Statistics
Nolan/Speed: Stat Labs: Mathematical Statistics Through Applications
Pitman: Probability
Rawlings/Pantula/Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Second Edition

(continued after index)

Jean-Michel Marin
Christian P. Robert

Bayesian Core: A Practical Approach to Computational Bayesian Statistics

Springer

Jean-Michel Marin
Projet Select
INRIA Futurs
Laboratoire de Mathématiques
Université Paris–Sud
91405 Orsay Cedex
France
jean-michel.marin@math.u-psud.fr

Christian P. Robert
CREST-INSEE
and
CEREMADE
Université Paris–Dauphine
75775 Paris Cedex 16
France
xian@ceremade.dauphine.fr

Cover illustration: Artwork of Michel Marin, entitled *Pierre de Rosette*.

Library of Congress Control Number: 2006932972

ISBN-10: 0-387-38979-2 e-ISBN-10: 0-387-38983-0
ISBN-13: 978-0-387-38979-0 e-ISBN-13: 978-0-387-38983-7

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To our most challenging case studies,
Lucas, Joachim and Rachel

Preface

After that, it was down to attitude.
—**Ian Rankin**, ***Black & Blue***.—

The purpose of this book is to provide a self-contained (we insist!) entry into practical and computational Bayesian statistics using generic examples from the most common models for a class duration of about seven blocks that roughly correspond to 13 to 15 weeks of teaching (with three hours of lectures per week), depending on the intended level and the prerequisites imposed on the students. (That estimate does not include practice—i.e., programming labs—since those may have a variable duration, also depending on the students' involvement and their programming abilities.) The emphasis on *practice* is a strong feature of this book in that its primary audience consists of graduate students who need to use (Bayesian) statistics as a tool to analyze their experiments and/or datasets. The book should also appeal to scientists in all fields, given the versatility of the Bayesian tools. It can also be used for a more classical statistics audience when aimed at teaching a quick entry to Bayesian statistics at the end of an undergraduate program for instance. (Obviously, it can supplement another textbook on data analysis at the graduate level.)

The format of the book is of a rather sketchy coverage of the topics, always backed by a motivated problem and a corresponding dataset (available on the Website of the course), and a detailed resolution of the inference procedures pertaining to this problem, sometimes including commented R programs. Special attention is paid to the derivation of prior distributions, and operational reference solutions are proposed for each model under study. Additional cases are proposed as exercises. The spirit is not unrelated to that of

Nolan and Speed (2000), with more emphasis on the theoretical and methodological backgrounds. We originally planned a complete set of lab reports, but this format would have forced us both to cut on the methodological side and to increase the description of the datasets and the motivations for their analysis. The current format is therefore more self-contained (than it would have been in the lab scenario) and can thus serve as a unique textbook for a service course for scientists aimed at analyzing data the Bayesian way or as an introductory course on Bayesian statistics.

A course corresponding to the book has now been taught by both of us for three years in a second year master's program for students aiming at a professional degree in data processing and statistics (at Université Paris Dauphine, France). The first half of the book was used in a seven-week (intensive) program, and students were tested on both the exercises (meaning all exercises) and their (practical) mastery of the datasets, the stated expectation being that they should go beyond a mere reproduction of the R outputs presented in the book. While the students found that the amount of work required by this course was rather beyond their usual standards (!), we observed that their understanding and mastery of Bayesian techniques were much deeper and more ingrained than in the more formal courses their counterparts had in the years before. In short, they started to think about the purpose of a Bayesian statistical analysis rather than on the contents of the final test and they ended up building a true intuition about what the results should look like, intuition that, for instance, helped them to detect modeling and programming errors! In most subjects, working on Bayesian statistics from this perspective created a genuine interest in the approach and several students continued to use this approach in later courses or, even better, on the job.

Contrary to usual practice, the exercises are interspersed within the chapters rather than postponed until the end of each chapter. There are two reasons for this stylistic choice: First, the results or developments contained in those exercises are often relevant for upcoming points in the chapter. Second, they signal to the student (or to any reader) that some pondering over the previous pages may be useful before moving to the following topic and so may act as self-checking gateways.

Thanks

We are immensely grateful to colleagues and friends for their help with this book, in particular, to the following people: François Perron somehow started us thinking about this book and did a thorough editing of it during a second visit to Dauphine, helping us to adapt it more closely to North American audiences. Charles Bouveyron provided and explained the vision dataset of Chapter 8. Jean-François Cardoso provided the cosmological background data in Chapter 2. George Casella made helpful suggestions on the format of the book. Gilles Celeux carefully read the manuscript and made numerous suggestions on both content and style. Noel Cressie insisted on a spatial chapter in

the "next" book (even though Chapter 8 may not be what he had in mind!). Jérôme Dupuis provided capture-recapture slides that have been recycled in Chapter 5. Arnaud Doucet and Chris Holmes made helpful suggestions during a memorable dinner in Singapore (and later Arnaud used a draft of the book in his class at the University of British Columbia, Vancouver). Jean-Dominique Lebreton provided the European dipper dataset of Chapter 5. Gaelle Lefol pointed out the Eurostoxx series as a versatile dataset for Chapter 7. Kerrie Mengersen collaborated with both of us on a review paper about mixtures that is related to Chapter 6 (and also gave us plenty of information about a QTL dataset that we ended up not using). Jim Kay introduced us to the Lake of Menteith dataset. Mike Titterington is thanked for collaborative friendship over the years and for a detailed set of comments on the book (quite in tune with his dedicated editorship of *Biometrika*). We are also grateful to John Kimmel of Springer for his advice and efficiency, as well as to two anonymous referees. Students and faculty members who attended the Finish MCMC spring 2004 course in Oulanka also deserve thanks both for their dedication and hard work, and for paving the ground for this book. In particular, the short introduction to R in Chapter 1 is derived from a set of notes written for this spring course. Teaching the highly motivated graduate students of Universidad Carlos III, Madrid, a year later, also convinced the second author that this venture was realistic. Later invitations to teach from this book both at the University of Canterbury, Christchurch (New Zealand) and at the Universidad Central de Venezuela, Caracas (Venezuela), were welcome indicators of its appeal, for which we are grateful to both Dominic Lee and José Léon. In addition, Dominic Lee and the students of STAT361 at the University of Canterbury very timely pointed out typos and imprecisions that were taken into account before the manuscript left for the printer.

This book was written while the first author was on leave as a Chargé de Recherche in the Unité Futurs of the Institut National de la Recherche en Informatique et Automatique (INRIA). He is grateful to both INRIA Futurs and the Université Paris Dauphine for granting him the time necessary to work on this project in the best possible conditions. The first author is, in addition, grateful to all his colleagues at the Université Paris Dauphine for their constant support and to Gilles Celeux from INRIA Futurs for his warm welcome and his enthusiastic collaboration in the Select project. *Enfin, le premier auteur salue toutes les personnes sans lesquelles cet ouvrage n'aurait jamais vu le jour ; d'un point de vue scientifique, il pense notamment à Henri Caussinus et Thierry Dhorne; d'un point de vue personnel, il remercie Carole Bégué pour son inestimable support, il pense aussi à Anne-Marie Dalas et Anne Marin.* Both authors are also grateful to Michel Marin, who designed the cover of the book.

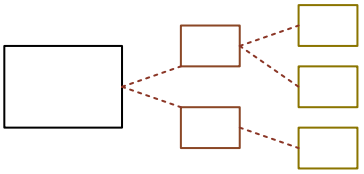
Parts of this book were also written on trips taken during the sabbatical leave of the second author: He is grateful to the Comité National des Universités (CNU) for granting him this leave, and, correlatively, to both the Department of Statistics, University of Glasgow, Scotland (hence, the Rankin quotations!), and the Institute for Mathematical Sciences, National University of Singapore, for their invaluable hospitality. He is also indebted to the University of Canterbury, Christchurch (New Zealand), for granting him a Visiting Erskine Fellowship in 2006 to teach out of this book. Special thanks, too, go to Hotel Altiplanico, San Pedro de Atacama (Chile), for providing albeit too briefly the ultimate serene working environment! The second author has an even more special thought for Bernhard K.N.T. Flury, whose data we use in Chapter 4 and who left us in the summer of 1998 for a never-ending climb of *ætheral via ferratas*. *Et, pour finir, des mercis très spéciaux à Brigitte, Joachim et Rachel pour avoir été là, à Denis pour les fractionnés du mercredi midi, et à Baptiste pour ses relais parfois vitaux !*

Paris **Jean-Michel Marin**
December 15, 2006 **Christian P. Robert**

Contents

1

User's Manual



The bare essentials, in other words.
—**Ian Rankin**, ***Tooth & Nail***.—

Roadmap

The **Roadmap** is a section that will start each chapter by providing a commented table of contents. It also usually contains indications on the purpose of the chapter.

For instance, in this first chapter, we explain the typographical notations that we adopted to distinguish between the different semantic levels of the course. We also try to detail how one should work with this book and how one could best benefit from this work. This chapter is to be understood as a user's (or instructor's) manual that details our pedagogical choices. It also seems the right place to briefly introduce the programming language R, which we use to analyze our datasets. (Note that it is not necessary to understand this language to benefit from the book!) For an introduction to Bayesian statistics and MCMC methods, one has to wait at least until the next chapter.

In each chapter, both Ian Rankin's quotation and the figure on top of the title page are (at best) vaguely related to the topic of the chapter, and one should not waste too much time pondering their implications and multiple meanings.

1.1 Expectations

The key word associated with this book is *modeling*, that is, the ability to build up a probabilistic interpretation of an observed phenomenon and the "story" that goes with it. The "grand scheme" is to get anyone involved in analyzing data to process a dataset within this coherent methodology. This means picking a parameterized probability distribution, denoted by f_θ, and extracting information about (shortened in "estimating") the unknown parameter θ of this probability distribution in order to provide a convincing interpretation of the reasons that led to the phenomenon at the basis of the dataset (and/or to be able to draw predictions about upcoming phenomena of the same nature). Before starting the description of the probability distributions, we want to impose on the reader the essential feature that a model is an *interpretation* of a real phenomenon that fits its characteristics to some degree of approximation rather than an *explanation* that would require the model to be "true." In short, there is no such thing as a "true model," even though some models are more appropriate than others!

In this book, we chose to describe the use of "classical" probability models for several reasons: First, it is often better to start a trip on well-traveled paths because they are less likely to give rise to unexpected surprises and misinterpretations. Second, they can serve as references for more advanced modelings: Quantities that appear in both simple and advanced modelings should get comparable estimators or, if not, the more advanced modeling should account for that difference. At last, the deliberate choice of an artificial model should give a clearer meaning to the motto that *all models are false* in that it illustrates the fact that a model is not necessarily justified by the theory beyond the modeled phenomenon but that its corresponding inference can nonetheless be exploited *as if* it were a true model.[1] By the end of the book, the reader should also be in a position to assess the relevance of a particular model for a given dataset.

Working with this book should not appear as a major endeavor: The datasets are described along with the methods that are relevant for the corresponding model, and the statistical analysis is provided with detailed comments. (R programs that support this analysis also are available.) If there is a difficulty, it actually starts at this point: Once the reader has seen the analysis, it should be possible to repeat this analysis or a similar analysis without further assistance. Even better, she or he should try to read as little as possible of the analysis proposed in this book and on the opposite hand should try to conduct the following stage of the analysis *before* reading the proposed (but not unique) solution. The ultimate lesson here is that there are indeed many ways to analyze a dataset and to propose modeling scenarios and inferential schemes. It is beyond the purpose of this book to provide all of those analyses,

[1] Working with more complex models often leads to the illusion that if they are complex enough they must be able to fit the "true" model!

and the reader (or the instructor) is supposed to look for alternatives on her or his own.

We thus expect readers to place themselves in a realistic situation to conduct this analysis in life-threatening (or job-threatening) situations. As detailed in the preface, the course was originally intended for students in the last year of study toward a professional degree, and it seems quite reasonable to insist that they face similar situations before entering their incoming job!

All datasets and R programs are available on the Website associated with the book. They should thus be used freely by the students in preparation for the following class as well as while studying the previous class.

1.2 Prerequisites and Further Reading

This being a textbook about statistical modeling, the students are supposed to have a background in both probability and statistics, at the level, for instance, of Casella and Berger (2001). In particular, a knowledge of standard sampling distributions and their properties is desirable. Lab work in the spirit of Nolan and Speed (2000) is also a plus. (One should read in particular their Appendix A on "How to write lab reports?") Further knowledge about Bayesian statistics is not a requirement, although using Robert (2001) or Lee (1997) as further references will bring a better insight into the topics treated here.

Exercise 1.1. Given a function g on $\mathbb{R}$, state the two basic conditions for g to be a probability density function (pdf) with respect to the Lebesgue measure.[2] Recall the definition of the cumulative distribution function (cdf) associated with g and that of the quantile function of g.

Exercise 1.2. If (x_1, x_2) is a normal $\mathcal{N}_2((\mu_1, \mu_2), \Sigma)$ random vector, with

$$\Sigma = \begin{pmatrix} \sigma^2 & \omega\sigma\tau \\ \omega\sigma\tau & \tau^2 \end{pmatrix},$$

recall the conditions on (ω, σ, τ) for Σ to be a (nonsingular) covariance matrix. Under those conditions, derive the conditional distribution of x_2 given x_1.

Similarly, we expect students to be able to understand the bits of R programs provided in the analysis, mostly because the syntax of R is very simple. We include a short and optional introduction to this language in the final

[2] In this book, the pdfs of continuous random variables are always defined with respect to the Lebesgue measure and those of discrete random variables (i.e., of random variables defined on a countable space) are defined in terms of the counting measure.

section of this chapter and we refer to Dalgaard (2002) for a deeper entry and also to Venables and Ripley (2002). Obviously, readers who are proficient with another computer language are free to experiment with this other language: The statistical output is truly what matters and, while we chose R for several good reasons (freeware, open source, good graphical and statistical abilities, accessibility), readers must come to the realization that a statistical practitioner will have to move from one language to another when changing employers or even positions. A short primer on R is provided at the end of this chapter.

Besides Robert (2001), the philosophy of which is obviously reflected in this book, other reference books pertaining to applied Bayesian statistics include Gelman et al. (2001), Carlin and Louis (1996), and Congdon (2001, 2003). More specific books that cover parts of the topics of a given chapter will be mentioned (with moderation) in the corresponding chapter, but we can quote here the relevant books of Holmes et al. (2002), Pole et al. (1994), and Gill (2002). We want to stress that the citations are limited for efficiency purposes: There is no extensive coverage of the literature as in, e.g., Robert (2001) or Gelman et al. (2001), because the prime purpose of the book is to provide a working methodology, for which incremental improvements and historical perspectives are not directly relevant.

While we also cover simulation–based techniques in a self-contained perspective, and thus do not assume prior knowledge of Monte Carlo methods, detailed references are Robert and Casella (2004) and Chen et al. (2000).

1.3 Styles and Fonts

Presentation often matters almost as much as content, and this is particularly true for data analyzes, since they aim to reproduce a realistic situation of a consultancy job where the consultant must report to a customer the results of an analysis. An equilibrated use of graphics, tables, itemized comments, and short paragraphs is, for instance, quite important for providing an analysis that stresses the different conclusions of the work, as well as the points that are yet unclear and those that could be expanded.

In particular, because this book is trying to do several things at once (that is, to introduce theoretical and computational concepts and to implement them in realistic situations), it needs to differentiate between the purposes and the levels of the parts of the text so that it is as obvious as possible to the reader. To this effect, we take advantage of the many possibilities of modern computer editing, and in particular of LaTeX, as follows.

First, a minimal amount of theoretical bases is required for dealing with the model introduced in each chapter, either for Bayesian statistics or for Monte Carlo theory. This aspect of the material will be part of the main text, but it will be kept to a minimum—just enough for the book to be self-contained—and therefore references to more detailed books such as Robert (2001) and

Robert and Casella (2004) will be necessary. These sections will need to be well-understood before handling the following applications or realistic cases. This book is primarily intended for those without a strong background in the theory of Bayesian statistics or computational methods, and "theoretical" sections are essential for them, hence the need to keep those sections within the main text.[3]

Ⓒ Although the whole book is intended to provide practical training in the Bayesian analysis of parameterized statistical models, some sections will be more practically oriented than others, especially in terms of computer implementation, and they will be signaled as the current paragraph with a Ⓒ symbol in the margin and the *sans sérif* font. These computing details will be separated from the theory of computation, represented in the main style.

Algorithms are often used in this book and are produced in code-like style (using `while` or `for` loops and `if` statements) or, more rarely, in the R language, although, again, no computer language is assumed known as a preliminary to the course. These algorithms will be presented in boxes and numbered, so they are clearly identifiable (see, e.g., page 36).

Once again, more advanced references such as Gentle (2002) are available for a deeper understanding of computational issues.

⊞ Statistics is as much about data processing as about mathematical and probabilistic modeling. To enforce this principle, we will center each chapter around one or two specific realistic datasets that will be described early enough in the chapter to be used extensively throughout the chapter. These datasets will be available on the book's Website[4] as **normaldata**, **capturedata**, and so on, the name being chosen in reference to the case/chapter heading. (Some of these datasets are already available as `datasets` in the R language.) In particular, we will explain the "how and why" of the corresponding dataset in a separate paragraph with *sans sérif* fonts and a window ⊞ in the margin (as in "window on the world" rather than as in a well-known registered[5] software trademark!). This style will also be used for illustrations of theoretical developments for the corresponding dataset and for specific computations related to this dataset. For typographical convenience, large graphs and tables may appear outside these sections, in subsequent pages, but will obviously be mentioned within them.

Example 1.1. There may also be a need for examples in addition to the main datasets, although we strived to keep them to a minimum and only for very specific issues where the reference dataset was not appropriate. They follow this numbered style, the sideways triangle indicating the end of the example. ◀

[3] We also emphasize the point that this book was first written for (and tested on) a graduate French audience and thus that their theoretical expectations may differ from those of other audiences: For a standard graduate French audience, the current approach is very, very evasive about theoretical details!

[4] At the present time, there exists a version of this Website on the authors' homepages, as well as a mirror at the University of Canterbury (NZ).

[5] Actually, this book was entirely produced, simulations and computations included, using a Linux system with GNU software.

Less important parts of the text such as extensions, side remarks, links to other topics, technical details, exercises, etc., will be separated from the main text in this shaded format. Although exercises are an essential part of the book, they are associated with this style because they can possibly be skipped during a first reading of the corresponding section (and also during a lecture). Note, however, that the results of some exercises will be used later without additional details.

The last style used in the book is the warning, represented by a lightning symbol in the margin: This entry is intended to signal major warnings about things that can (and do) go wrong "otherwise"; that is, if the warning is not taken into account. Needless to say, these paragraphs must be given the utmost attention!

1.4 A Short Introduction to R

Let us start with two preliminary statements: First, as already indicated above, the book does not require knowledge of the R programming language since there are many alternative ways of implementing Bayesian principles into computer programs.[6] Second, this section is not a proper introduction to R in the sense that additional efforts will be required from the reader before using the language. However, the syntax of R is simple enough to allow learning by trial-and-error in a few hours.

The main reason why we advocate using R as the programming interface between Bayesian methodology and the analysis of real data is that it combines a remarkable power (for an interpreted language) with a very clear syntax both for statistical computation and graphics. It also offers the advantages of being a free and open–source system, with constant upgrades and improvements as well as numerous Web-based tutorials and user's manuals, and of running on all platforms, including Apple and Windows (and, obviously, on Linux and Unix). The package is straightforward to install: It can be downloaded from one of the CRAN (Comprehensive R Archive Network) Websites.[7] The online help commands `help()` and `help.search()` are very good starting points to gather information about a specific function or a general issue.

[6] For instance, a ubiquitous programming language specially designed for Bayesian statistics is WinBUGS, developed by the Medical Research Council Unit of Cambridge University.

[7] The main CRAN Website is http://cran.r-project.org/+.

1.4.1 R Objects

The basic elements of the R language are the *objects*. An R object can be of many types, including vector, matrix, time series, data frames, functions, or graphics. It is mostly characterized by a *mode* that describes its contents and a *class* that describes its structure. The different modes are null (empty object), logical, numeric, complex, and character and the main classes are vector, matrix, array, factor, time-series, data.frame, and list. Heterogeneous objects such as those of the list class can include elements with various modes.

Since R is most efficient as a vectorial (or matricial) language,[8] let us start with the vector class. As indicated by its name, it is a vector of elements of the same type, such as `(TRUE,TRUE,FALSE,TRUE)` or `(1,2,3,5,7,11)`. Creating small vectors can be done using the R command `c()`. This function combines or concatenates terms together. Note that decimal numbers should be encoded with a dot, character strings in inverted commas `" "`, and logical values with the character strings `TRUE` and `FALSE` or their respective abbreviations `T` and `F`. Missing values are encoded with the character string `NA`.

In Figure 1.1, we give a few illustrations of the use of vectors in R. The symbol `>` at the beginning of each line is called the prompt and precedes the line command, which is terminated by pressing `RETURN`. The character + indicates that the console is waiting for a supplementary instruction, which is useful when typing long expressions. The assignment operator is `=`, not to be confused with `==`, which is the Boolean operator for equality. Note that new objects are simply defined by assigning a value, as in the first line of Figure 1.1.

Exercise 1.3. Test the `help()` command on the functions `seq()`, `sample()`, and `order()`. (*Hint:* Start with `help()`.)

The matrix class provides the R representation of matrices. For instance, `matrix(vec,nrow=n,ncol=p)` is an $n \times p$ matrix whose elements are those of `vec`, assuming this vector is of dimension np, stored by column except if the option `byrow=T` is used. The matricial product is denoted by `%*%`, while `*` represents the term-by-term product. (Note that taking the product `a%*%b` when the number of columns of `a` differs from the number of rows of `b` induces an error message.) Figure 1.2 gives a few examples of matrix–related commands. The `apply()` function is particularly efficient for applying functions on matrices by row or by column.

Structures with more than two indices are represented by *arrays* and can also be processed by R commands, as for instance `x=array(1:50,c(2,5,5))`, which gives a three-entry table of 50 terms.

[8] In the sense that avoiding loops always brings an improvement in execution time.

`> a=c(5,5.6,1,4,-5)`	build the object `a` containing a numeric vector of dimension 5 with elements 5, 5.6, 1, 4, –5
`> a[1]`	display the first element of `a`
`> b=a[2:4]`	build the numeric vector `b` of dimension 3 with elements 5.6, 1, 4
`> d=a[c(1,3,5)]`	build the numeric vector `d` of dimension 3 with elements 5, 1, –5
`> 2*a`	multiply each element of `a` by two and display the result
`> e=3/d`	build the numeric vector `e` of dimension 3 and elements 3/5, 3, –3/5
`> log(d*e)`	multiply the vectors `d` and `e` term by term and transform each term into its logarithm
`> sum(d)`	calculate the sum of `d`
`> length(d)`	display the length of `d`
`> t(d)`	transpose `d`, the result is a row vector
`> t(d)%*%e`	scalar product between the row vector `t(b)` and the column vector `e`
`> g=c(sqrt(2),log(10))`	build the numeric vector `g` of dimension 2 and elements $\sqrt{2}$, log(3)
`> e[d==5]`	build the subvector of `e` that contains the components `e[i]` such that `d[i]=5`
`> is.vector(d)`	display the logical expression `TRUE` if a vector and `FALSE` else

Fig. 1.1. Illustrations of the processing of vectors in R.

`> b=sample(1:10,20,rep=T)`	produce a vector of 10 terms uniformly distributed on $\{1,\ldots,10\}$
`> x1=matrix(1:20,nrow=5)`	build the numeric matrix `x1` of dimension 5×4 with first row 1, 6, 11, 16
`> x2=matrix(1:20,nrow=5,byrow=T)`	build the numeric matrix `x2` of dimension 5×4 with first row 1, 2, 3, 4
`> x3=t(x2)`	transpose the matrix `x2`
`> b=x3%*%x2`	matrix product between `x2` and `x3`, including a check of the dimension
`> dim(x1)`	display the dimension of `x1`
`> b[,2]`	select the second column of `b`
`> b[c(3,4),]`	select the third and fourth rows of `b`
`> b[-2,]`	delete the second row of `b`
`> a[a<5]=0`	set to zero elements of `a` less than 5, but note that `a[a<5]` is a vector
`> rbind(x1,x2)`	vertical merging of `x1` and `x2`
`> apply(x1,1,sum)`	calculate the sum of each row of `x1`

Fig. 1.2. Illustrations of the processing of matrices in R.

A factor is a vector of characters or integers used to specify a discrete classification of the components of other vectors of the same length. The factor

is used to represent qualitative variables. R provides both ordered and unordered factors, whose major appeal lies within model formulas, as illustrated in Figure 1.3. Note the subtle difference between `apply()` and `tapply()`.

A list in R is a rather loose object made of a collection of other arbitrary objects known as its *components*. For instance, a list can be derived from n existing objects as follows:

```
a=list(name_1=object_1,...,name_n=object_n).
```

This function creates a list with n arguments using `object_1,...,object_n` for the components, each being associated with the argument's name, `name_i`. For instance, `a$name_1` will be equal to `object_1`. (It can also be represented as `a[[1]]` but this is less practical.) Lists are very useful in preserving information about the values of variables used within R functions, in the sense that all relevant values can be put within a list that is the output of the corresponding function (see Section 1.4.3 for details about the construction of functions in R). Note the use of the abbreviations `vec` and `val` in the last line of Figure 1.4: They are acceptable as long as they do not induce confusion.

A last class we just mention is the data frame. A data frame is a list whose elements are possibly made of differing modes and attributes but have the same length, as in the example provided in Figure 1.5. A data frame can be displayed in matrix form, and its rows and columns can be extracted using matrix indexing conventions. A list whose components satisfy the restrictions imposed on a data frame can be coerced into a data frame using the function `as.data.frame()`. The main purpose of this object is to import data from an external file by using the `read.table()` function.

```
> state=c("tas","tas","sa","sa","wa") create a vector with three values
> statef=factor(state)                distinguish entries by group
> levels(statef)                      give the groups
> incomes=c(60,59,40,42,23)
> tapply(incomes,statef,mean)         average the incomes for each group
```

Fig. 1.3. Illustrations of the factor class.

```
> li=list(num=1:5,y="color",a=T)
> a=matrix(c(6,2,0,2,6,0,0,0,36),nrow=3)
> res=eigen(a,symmetric=T)                diagonalize a and
> names(res)                              produce a list with two
> res$vectors                             arguments: vectors and values
> diag(res$values)                        create the diagonal matrix
                                          of eigenvalues
> res$vec%*%diag(res$val)%*%t(res$vec)    recover a
```

Fig. 1.4. Chosen features of the list class.

```
> v1=sample(1:12,30,rep=T)
```
simulate 30 independent uniform random variables on $\{1, 2, \ldots, 12\}$

```
> v2=sample(LETTERS[1:10],30,rep=T)
```
simulate 30 independent uniform random variables on $\{a, b, ..., j\}$

```
> v3=runif(30)
```
simulate 30 independent uniform random variables on $[0, 1]$

```
> v4=rnorm(30)
```
simulate 30 independent realizations from a standard normal distribution

```
> xx=data.frame(v1,v2,v3,v4)
```

Fig. 1.5. A definition of a data frame.

1.4.2 Probability Distributions in R

R is primarily a statistical language. It is therefore well-equipped with probability distributions. As described in Table 1.1, all standard distributions are available, with a clever programming shortcut: A "core" name, such as `norm`, is associated with each distribution and the four basic associated functions, namely the cdf, the pdf, the quantile function, and the simulation procedure, are defined by appending the prefixes `d, p, q, r` to the core name, such as `dnorm()`, `pnorm()`, `qnorm()`, and `rnorm()`. Obviously, each function requires additional entries, as in `pnorm(1.96)` or `rnorm(10,mean=3,sd=3)`. Recall that `pnorm()` and `qnorm()` are inverses of one another.

Table 1.1. Standard distributions with R core name.

Distribution	Core	Parameters	Default Values
Beta	`beta`	`shape1, shape2`	
Binomial	`binom`	`size, prob`	
Cauchy	`cauchy`	`location, scale`	0, 1
Chi-square	`chisq`	`df`	
Exponential	`exp`	`1/mean`	1
Fisher	`f`	`df1, df2`	
Gamma	`gamma`	`shape,1/scale`	`NA`, 1
Geometric	`geom`	`prob`	
Hypergeometric	`hyper`	`m, n, k`	
Log-Normal	`lnorm`	`mean, sd`	0, 1
Logistic	`logis`	`location, scale`	0, 1
Normal	`norm`	`mean, sd`	0, 1
Poisson	`pois`	`lambda`	
Student	`t`	`df`	
Uniform	`unif`	`min, max`	0, 1
Weibull	`weibull`	`shape`	

In addition to these probability functions, R also provides a battery of (classical) statistical tools, ranging from descriptive statistics to nonparametric tests and generalized linear models. A description of these abilities is not possible in this section but we refer the reader to, e.g., Dalgaard (2002) or Venables and Ripley (2002) for a complete entry.

Exercise 1.4. Study the properties of the R function `lm()` using simulated data as in

```
> x=rnorm(20)
> y=3*x+5+rnorm(20,sd=0.3)
> reslm=lm(y~x)
> summary(reslm)
```

Another appeal of the R language is its range of graphical possibilities. Functions such as `plot()` and `image()` can be customized to a large extent, as described in Venables and Ripley (2002). (For instance, colors can be chosen by `colors()` out of 650 hues, and LaTeX–like formulas can be included within the graphs using `expression`, such as `expression(mu[i])` for μ_i.) Figure 1.6 gives some of the most usual graphical commands.

1.4.3 Writing New R Functions

One of the strengths of R is that new functions and libraries can be created by anyone and then added to Web depositories to continuously enrich the language. These new functions are not distinguishable from the core functions of R, such as `median()` or `var()`, because those are also written in R. This means their code can be accessed and potentially modified, although it is safer to define new functions. (A few functions are written in C, though, for efficiency.) Learning how to write functions designed for one's own problems

```
> x=rnorm(100)
> hist(x,nclass=10, prob=T)                    compute and plot an histogram
                                               of x
> curve(dnorm(x),add=T)                        draw the normal density on top
> y=2*x+rnorm(100,0,2)
> plot(x,y,xlim=c(-5,5),ylim=c(-10,10))        draw a scatterplot of x against y
> lines(c(0,0),c(1,2),col="sienna3")
> boxplot(x)                                   compute and plot
                                               a box-and-whiskers plot of x
> state=c("tas","tas","sa","sa","wa","sa")
> statef=factor(state)
> barplot(table(statef))                       draw a bar diagram of x
```

Fig. 1.6. Some standard plotting commands.

is paramount for their resolution, even though the huge collection of available R functions may often contain a function already written for that purpose.

Exercise 1.5. Of the R functions you have met so far, check which ones are written in R by simply typing their name without parentheses, as in `mean` or `var`.

A function is defined in R by an assignment of the form

```
name=function(arg1[=expr1],arg2[=expr2],...) {
expression
...
expression
value
}
```

where `expression` denotes an R command that uses some of the arguments `arg1, arg2, ...` to calculate a value, `value`, that is the outcome of the function. The braces indicate the beginning and the end of the function and the brackets some possible default values for the arguments. Note that producing a value at the end of a function is essential because anything done within a function is local and temporary, and therefore lost once the function has been exited, unless saved in `value` (hence, again, the appeal of `list()`). For instance, the following function, named `titus`, implements a version of Newton's method for calculating the square root of `y`:

```
titus=function(y) {
x=y/2
while (abs(x*x-y) > 1e-10) x=(x+y/x)/2
x
}
```

When designing a new R function, it is more convenient to use an external text editor and to store the function under development in an external file, say `myfunction.R`, which can be executed in R as `source("myfunction.R")`. Note also that some external commands can be launched within an R function via the very handy command `system()`. This is, for instance, the easiest way to incorporate programs written in other languages (e.g., Fortran, C, Matlab) within R programs.

The expressions used in a function rely on a syntax that is quite similar to those of other programming languages, with conditional statements such as

```
if (expres1) expres2 else expres3
```

where `expres1` is a logical value, and loops such as

```
for (name in expres1) expres2
```

and

```
while (name in expres1) expres2
```

where `expres1` is a collection of values, as illustrated in Figure 1.7. In particular, Boolean operators can be used within those expressions, including `==` for testing equality, `!=` for testing inequality, `&` for the logical and, `|` for the logical or, and `!` for the logical contradiction.

```
> bool=T;i=0                                separate commands by semicolons
> while(bool==T) {i=i+1; bool=(i>10)}       stop at i = 11
> s=0;x=rnorm(10000)
> system.time(for (i in 1:length(x)){       output sum(x) and
+    s=s+x[i]})[3]                          provide computing time
> system.time(t(rep(1,10000))%*%x)[3]       compare with vector product
> system.time(sum(x))[3]                    compare with sum() efficiency
```

Fig. 1.7. Some artificial loops in R.

Since R is an interpreted language, avoiding loops is generally a good idea, but this may render programs much harder to read. It is therefore extremely useful to include comments within the programs by using the symbol `#`.

1.4.4 Input and Output in R

Large data objects need to be read as values from external files rather than entered during an R session at the keyboard (or by cut-and-paste). Input facilities are simple, but their requirements are fairly strict. In fact, there is a clear presumption that it is possible to modify input files using other tools outside R.

An entire data frame can be read directly with the `read.table()` function. Plain files containing rows of values with a single mode can be downloaded using the `scan()` function, as in

```
> a=matrix(scan("myfile"),nrow=5,byrow=T)
```

When data frames have been produced by another statistical software, the library `foreign` can be used to input those frames in R. For example, the function `read.spss()` allows ones to read SPSS data frames.

Conversely, the generic function `save()` can be used to store all R objects in a given file, either in binary or ASCII format. (The alternative function `dump()` is more rudimentary but also useful.) The function `write.table()` is used to export R data frames as ASCII files.

1.4.5 Administration of R Objects

During an R session, objects are created and stored by name. The command `objects()` (or, alternatively, `ls()`) can be used to display, within a directory called the *workspace*, the names of the objects that are currently stored. Individual objects can be deleted with the function `rm()`.

All objects created during an R session (including functions) can be stored permanently in a file in provision of future R sessions. At the end of each R session, obtained by the command `quit()` (which can be abbreviated as `q()`), the user is given the opportunity to save all the currently available objects, as in

```
>q()
Save workspace image? [y/n/c]:
```

If the user answers `y`, the object created during the current session and those saved from earlier sessions are saved in a file called `.RData` and located in the working directory. When R is called again, it reloads the workspace from this file, which means that the user starts the new session exactly where the old one had stopped. In addition, the entire past command history is stored in the file `.Rhistory` and can be used in the current or in later sessions by using the command `history()`.

Note that launching an R session automatically loads the function contained in the base R library `base`. Functions contained in other libraries (whose list is available via `library()`) must be loaded manually by the user, as in `library(stats)`.

2

Normal Models

This was where the work really took place.
—**Ian Rankin**, ***Knots & Crosses***.—

Roadmap

As in every subsequent chapter, we start with a description of the data used for the whole chapter as a benchmark for illustrating new methods and for testing assimilation of the techniques. We then propose a corresponding statistical model centered on the normal $\mathcal{N}(\mu, \sigma^2)$ distribution and consider specific inferential questions to address at this level, namely parameter estimation, one-sided test, prediction, and outlier detection, after we set the description of the Bayesian resolution of inferential problems. This being the first chapter, the amount of technical/theoretical material may be a little overwhelming at times. It is, however, necessary to go through these preliminaries before getting to more advanced topics with a minimal number of casualties!

Exercise 2.1. Check your current knowledge of the normal $\mathcal{N}(\mu, \sigma^2)$ distribution by writing down its density function and computing its first four moments.

Exercise 2.2. Before exiting to the next page, think of datasets that could be, or could not be, classified as normal. In each case, describe your reason for proposing or rejecting a normal modeling.

2.1 Normal Modeling

The normal (or Gaussian) distribution $\mathcal{N}(\mu, \sigma^2)$, with density on $\mathbb{R}$,

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},$$

is certainly one of the most studied and one of the most used distributions because of its "normality": It appears both as the limit of additive small effects and as a representation of symmetric phenomena without long tails, and it offers many openings in terms of analytical properties and closed-form computations. As such, it is thus the natural opening to a modeling course, even more than discrete and apparently simpler models such as the binomial and Poisson models we will discuss in the following chapters. Note, however, that we do not advocate at this stage the use of the normal distribution as an all-purpose model: There exist many continuous situations where a normal univariate model is inappropriate for many possible reasons (e.g., skewness, fat tails, dependence, multimodality).

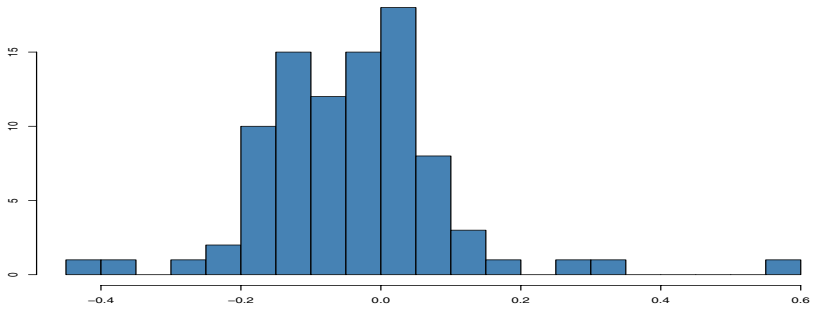


Fig. 2.1. Dataset normaldata: Histogram of the relative changes in reported larcenies between 1991 and 1995 in the 90 most populous US counties. (*Source:* US Bureau of Justice.)

⊞ Our first dataset, **normaldata**, describes the relative changes in reported larcenies between 1991 and 1995 (relative to 1991, that is, the difference is divided by the 1991 figures) for the 90 most populous US counties, as reported by the US Bureau of Justice and compiled from FBI sources. The 90 data points are represented in Figure 2.1 by an histogram that seems compatible with a symmetric unimodal distribution such as the normal distribution. The fit may not be perfect, though, because of (a) a possible bimodality of the histogram and (b) one outlier on the right at almost 0.6. While, in a regular setting, a normal $\mathcal{N}(\mu, \sigma^2)$ distribution with both parameters unknown would

have to be fitted to this dataset, for simplicity's sake we will start our illustration with a $\mathcal{N}(\mu, \hat{\sigma}^2)$ modeling, where $\hat{\sigma}^2$ is the empirical variance of the sample.

Exercise 2.3. Reproduce the histogram of Figure 2.1 and the subsequent analysis conducted in this chapter for the relative changes in reported larcenies relative to the 1995 figures using the file 90cntycr.wk1 available on the Webpage of this book.

As mentioned above, the use of a normal distribution for modeling a given dataset is a convenient device that does not need to correspond to a perfect fit. With some degree of approximation, the normal distribution may agree with the data sufficiently to be used in place of the true distribution (if any). There exist, however, some setups where the normal distribution is thought to be the exact distribution behind the dataset (or where departure from normality has a significance for the theory behind the observations). This is, for instance, the case with the following dataset, where normality is related to some finer details of the Big Bang theory of the origin of the universe.

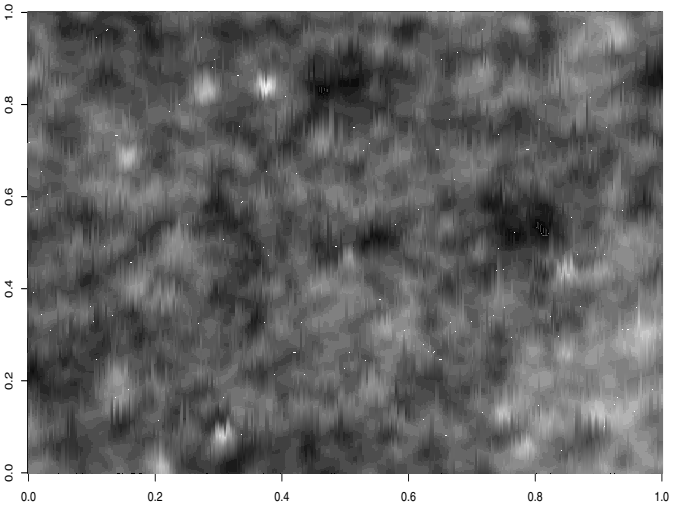


Fig. 2.2. Dataset CMBdata: Spectral image of the cosmological microwave background (CMB) of the universe. (The darker the pixel, the higher the temperature difference from the mean temperature.)

⊞ Figure 2.2 is an image (in the spectral domain) of the "cosmological microwave background" (CMB), in a region of the sky: More specifically, this picture represents the

electromagnetic radiation from photons dating back to the early ages of the universe, a radiation often called "fossil light," that dates back to a few hundred thousand years after the Big Bang (Chown, 1996).The grey levels are given by the differences in apparent temperature from the mean temperature and as stored in the **CMBdata**.

For astrophysical (or rather cosmological) reasons too involved to be detailed here, the repartition of the spectrum is quite isotropic (that is, independent of direction) and normal. In fact, if we treat each temperature difference in Figure 2.2 as an independent realization, the histogram of these differences, represented in Figure 2.3, provides a rather accurate representation of the distribution of these temperatures, along with a normal fit (based on the maximum likelihood estimate). This fit is actually extremely good since this histogram corresponds to the 800×800 = 640,000 points in Figure 2.2! In the overwhelming majority of cases, a normal fit to a dataset that size would see a much larger departure from normality on a portion of the dataset! (Actually, there also are departures from normality in the current case in that the brighter points on the image correspond to local lacks of homogeneity that are of interest for astrophysicists.)

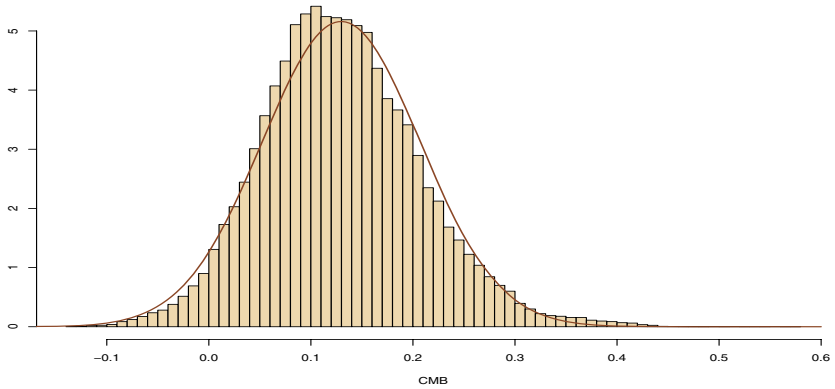


Fig. 2.3. Dataset CMBdata: Histogram and normal fit.

Exercise 2.4. By creating random subwindows of the region plotted in Figure 2.2, represent the histograms of these subsamples and examine whether they strongly differ or not. Pay attention to the possible influence of the few "bright spots" on the image.

2.2 The Bayesian Toolkit

2.2.1 Bases

Given an independent and identically distributed (iid) sample $\mathscr{D} = (x_1, \ldots, x_n)$ from a density f_θ, with an unknown parameter $\theta \in \Theta$, like the mean μ of the benchmark normal distribution, the associated likelihood function is

$$\ell(\theta|\mathscr{D}) = \prod_{i=1}^{n} f_\theta(x_i)\,. \tag{2.1}$$

This quantity is a fundamental entity for the analysis of the information provided about θ by the sample $\mathscr{D}$, and Bayesian analysis relies on this function to draw inference on θ. The major input of the Bayesian approach, compared with a standard likelihood approach, is that it modifies the likelihood into a *posterior* distribution, which is a probability distribution on Θ defined by

$$\pi(\theta|\mathscr{D}) = \frac{\ell(\theta|\mathscr{D})\pi(\theta)}{\int \ell(\theta|\mathscr{D})\pi(\theta)\,\mathrm{d}\theta}\,. \tag{2.2}$$

The factor $\pi(\theta)$ in (2.2) is called the *prior* and it obviously has to be determined to start the analysis. A first motivation for this approach is that the prior distribution summarizes the *prior information* on θ; that is, the knowledge that is available on θ *prior* to the observation of the sample $\mathscr{D}$. However, the choice of $\pi(\theta)$ is often decided on practical grounds rather than strong subjective beliefs or overwhelming prior information. As will be discussed later, there also exist less subjective choices, called *noninformative priors*.

Exercise 2.5. Show that (2.2) can be derived by first setting θ as a random variable with density function π and then $\mathscr{D}$ conditionally on θ as distributed from $\ell(\theta|\mathscr{D})$.

⊞ For the **normaldata** setting, we can assume that the relative change in reported larcenies has an average μ between -0.5 and 0.5 and, with no further prior information, we choose the prior to be the uniform $\mathscr{U}(-0.5, 0.5)$ distribution. The posterior is then given by

$$\pi(\mu|\mathscr{D}) = \frac{e^{-90(\mu-\overline{x})^2/2\sigma^2}}{\int_{-0.5}^{0.5} e^{-90(\theta-\overline{x})^2/2\sigma^2}\,\mathrm{d}\theta}\,\mathbb{I}_{[-0.5,0.5]}(\theta),$$

which is a truncated normal distribution.

The concept that is at the core of Bayesian analysis is that one should provide an inferential assessment *conditional on the realized value of* $\mathscr{D}$, and Bayesian analysis gives a proper probabilistic meaning to this conditioning by allocating to θ a probability distribution. Once the prior distribution is

selected, Bayesian inference formally is "over"; that is, it is completely determined since the estimation, testing, and evaluation procedures are automatically provided by the prior and the associated loss (or penalty) function. For instance, if estimations $\hat{\theta}$ of θ are compared via the quadratic loss function

$$\mathrm{L}(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2,$$

the corresponding Bayes procedure is the *expected* value of θ under the posterior distribution,[1]

$$\hat{\theta} = \int \theta \, \pi(\theta|\mathscr{D}) \, \mathrm{d}\theta = \frac{\int \theta \, \ell(\theta|\mathscr{D}) \, \pi(\theta) \, \mathrm{d}\theta}{\int \ell(\theta|\mathscr{D}) \, \pi(\theta) \, \mathrm{d}\theta}, \tag{2.3}$$

for a given sample $\mathscr{D}$.

Exercise 2.6. Show that the minimization (in $\hat{\theta}(\mathscr{D})$) of the expectation $\mathbb{E}[\mathrm{L}(\theta, \hat{\theta}))|\mathscr{D}]$—that is, of the expectation of the quadratic loss function under the distribution with density $\pi(\theta|\mathscr{D})$—produces the posterior expectation as the solution in $\hat{\theta}$.

When no specific loss function is available, the estimator (2.3) is often used as a default estimator, although alternatives are also available. For instance, the *maximum a posteriori estimator* (MAP) is defined as

$$\hat{\theta} = \arg\max_\theta \pi(\theta|\mathscr{D}) = \arg\max_\theta \pi(\theta)\ell(\theta|\mathscr{D}), \tag{2.4}$$

where the function to maximize is usually provided in closed form. However, numerical problems often make the optimization involved in finding the MAP far from trivial. Note also here the similarity of (2.4) with the maximum likelihood estimator (MLE): The influence of the prior distribution $\pi(\theta)$ progressively disappears with the number of observations, and the MAP estimator recovers the asymptotic properties of the MLE. See Schervish (1995) for more details on the asymptotics of Bayesian estimators.

2.2.2 Prior Distributions

The selection of the prior distribution is an important issue in Bayesian statistics. When prior information is available about the data or the model, it can (and must) be used in building the prior, and we will see some illustrations of this recommendation in the following chapters. In many situations, however,

[1] Estimators are functions of the data $\mathscr{D}$, while estimates are values taken by those functions. In most cases, we will denote them with a "hat" symbol, the dependence on $\mathscr{D}$ being implicit.

the selection of the prior distribution is quite delicate in the absence of reliable prior information, and generic solutions must be chosen instead. Since the choice of the prior distribution has a considerable influence on the resulting inference, this choice must be conducted with the utmost care. There actually exists a category of priors whose primary aim is to minimize the impact of the prior selection on the inference: They are called *noninformative* priors and we will detail them below.

When the model is from an exponential family of distributions (Robert, 2001, Section 3.3.3)—that is, a family of distributions with densities of the form

$$f_\theta(y) = h(y)\, \exp\{\theta \cdot R(y) - \Psi(\theta)\}\,, \qquad \theta, R(y) \in \mathbb{R}^p\,,$$

where $\theta \cdot R(y)$ denotes the canonical scalar product in $\mathbb{R}^p$—there exists a generic class of priors called the *class of conjugate priors*,

$$\pi(\theta|\xi,\lambda) \propto \exp\{\theta \cdot \xi - \lambda\Psi(\theta)\}\,,$$

which are parameterized by two quantities, $\lambda > 0$ and ξ, that are of the same nature as $R(y)$. The proportionality symbol $\propto$ is to be understood in terms of functions of θ.[2] These parameterized prior distributions on θ are thus such that the posterior distributions are of the same form; that is, they can be written as

$$\pi(\theta|\xi'(y),\lambda'(y))\,, \tag{2.5}$$

where $(\xi'(y),\lambda'(y))$ is defined in terms of the observation y (Exercise 2.9). Equation (2.5) simply says that the conjugate prior is such that the prior and posterior densities belong to the same parametric family of densities but with different parameters. Indeed, the parameters of the posterior density are "updated," using the observations, relative to the prior parameters. To avoid confusion, the parameters involved in the prior distribution on the model parameter are usually called *hyperparameters*. (They can themselves be associated with prior distributions, then called *hyperpriors*.)

Exercise 2.7. Show that the normal, binomial, geometric, Poisson, and exponential distributions are all exponential families.

Exercise 2.8. Show that, for an exponential family, $\Psi(\theta)$ is defined by the constraint that f_θ is a probability density and that the expectation of this distribution can be written as $\partial\Psi(\theta)/\partial\theta$, the vector of the derivatives of $\Psi(\theta)$ with respect to the components of θ.

[2]The relation $\pi(\theta|x) \propto \tilde{\pi}(\theta|x)$ means that the functions π and $\tilde{\pi}$ only differ by a multiplicative constant that may depend on x. This is not a contradictory statement in that the data x is fixed once observed! While being entirely rigorous, given that probability densities are uniquely determined by their functional form, computations using proportionality signs lead to greater efficiency in the derivation of posterior distributions.

Exercise 2.9. Show that the updated hyperparameters in (2.5) are given by

$$\xi'(y) = \xi + R(y)\,, \quad \lambda'(y) = \lambda + 1\,.$$

Find the corresponding expressions for $\pi(\theta|\xi, \lambda, y_1, \ldots, y_n)$.

The normal distribution $\mathscr{N}(\mu, 1)$ is a special case of an exponential family, with $\theta = \mu$, $R(x) = x$, and $\Psi(\mu) = \mu^2/2$. The corresponding conjugate prior is thus of the form $\pi(\mu|\xi, \lambda) \propto \exp\{\mu\xi - \lambda\mu^2/2\}$, which implies that the conjugate prior for the normal mean μ is also normal,

$$\mathscr{N}\left(\lambda^{-1}\xi, \lambda^{-1}\right)\,.$$

This means that, when choosing a conjugate prior in a normal setting, one has to select both a mean and a variance a priori. (In some sense, this is the advantage of using a conjugate prior, namely that one has to select only a few parameters to determine the prior distribution. Conversely, the drawback of conjugate priors is that the information known a priori on μ either may be insufficient to determine both parameters or may be incompatible with the structure imposed by conjugacy.)

Once ξ and λ are selected, the posterior distribution on μ is determined by Bayes' theorem,

$$\begin{aligned}\pi(\mu|x) &\propto \exp(x\mu - \mu^2/2)\, \exp(\xi\mu - \lambda\mu^2/2) \\ &\propto \exp\left\{-(1+\lambda)\left[\mu - (1+\lambda)^{-1}(x+\xi)\right]^2/2\right\}\,,\end{aligned}$$

which means that this posterior distribution is a normal distribution with mean $(1+\lambda)^{-1}(x+\xi)$ and variance $(1+\lambda)^{-1}$. An alternative representation of the posterior mean is

$$\frac{\lambda^{-1}}{1+\lambda^{-1}}\, x + \frac{1}{1+\lambda^{-1}}\, \lambda^{-1}\xi\,; \tag{2.6}$$

that is, a weighted average of the observation x and the prior mean $\lambda^{-1}\xi$. The smaller λ is, the closer the posterior mean is to x.

Exercise 2.10. Derive the posterior distribution for an iid sample $\mathscr{D} = (y_1, \ldots, y_n)$ from $\mathscr{N}(\theta, 1)$ and show that it only depends on the sufficient statistic $\overline{y} = \sum_{i=1}^n y_i/n$.

Exercise 2.11. Give the range of values of the posterior mean (2.6) as the pair $(\lambda, \lambda^{-1}\xi)$ varies over $\mathbb{R}^+ \times \mathbb{R}$.

As noted earlier, to use a normal distribution with known variance on a real dataset is quite unrealistic. We now consider the general case of an iid sample

$\mathscr{D} = (x_1, \ldots, x_n)$ from the normal distribution $\mathscr{N}(\mu, \sigma^2)$ and $\theta = (\mu, \sigma^2)$. This setting also allows a conjugate prior since the normal distribution remains an exponential family when both parameters are unknown. (This is not always the case; see Exercise 2.12.) It is of the form

$$(\sigma^2)^{-\lambda_\sigma} \exp\left\{-\left(\lambda_\mu(\mu-\xi)^2+\alpha\right)/2\sigma^2\right\}$$

since

$$\begin{aligned}\pi((\mu,\sigma^2)|\mathscr{D}) &\propto (\sigma^2)^{-\lambda_\sigma-3/2} \exp\left\{-\left(\lambda_\mu(\mu-\xi)^2+\alpha\right)/2\sigma^2\right\}\\ &\quad\times(\sigma^2)^{-n/2} \exp\left\{-\left(n(\mu-\overline{x})^2+s_x^2\right)/2\sigma^2\right\} \qquad (2.7)\\ &\propto (\sigma^2)^{-\lambda_\sigma(\mathscr{D})} \exp\left\{-\left(\lambda_\mu(\mathscr{D})(\mu-\xi(\mathscr{D}))^2+\alpha(\mathscr{D})\right)/2\sigma^2\right\}\,.\end{aligned}$$

Therefore, the conjugate prior on θ is the product of an inverse gamma distribution on σ^2, $\mathscr{IG}(\lambda_\sigma, \alpha/2)$, and, conditionally on σ^2, a normal distribution on μ, $\mathscr{N}(\xi, \sigma^2/\lambda_\mu)$.

Exercise 2.12. A Weibull distribution $\mathscr{W}(\alpha, \beta, \gamma)$ is defined as the power transform of a gamma $\mathscr{G}(\alpha, \beta)$ distribution: If $X \sim \mathscr{W}(\alpha, \beta, \gamma)$, then $X^\gamma \sim \mathscr{G}(\alpha, \beta)$. Show that, when γ is known, $\mathscr{W}(\alpha, \beta, \gamma)$ is an exponential family, but that it is not an exponential family when γ is unknown.

Exercise 2.13. Show that, when the prior on $\theta = (\mu, \sigma^2)$ is $\mathscr{N}(\xi, \sigma^2/\lambda_\mu) \times \mathscr{IG}(\lambda_\sigma, \alpha)$, the marginal prior on μ is a Student's t distribution $\mathscr{T}(2\lambda_\sigma, \xi, \alpha/\lambda_\mu\lambda_\sigma)$ (see Example 2.3 below for the definition of a Student's t density). Give the corresponding marginal prior on σ^2. For an iid sample $\mathscr{D} = (x_1, \ldots, x_n)$ from $\mathscr{N}(\mu, \sigma^2)$, derive the parameters of the posterior distribution of (μ, σ^2).

There is no compelling reason to choose *these* priors, except for their simplicity, but the restrictive aspect of conjugate priors can be attenuated by using *hyperpriors* on the hyperparameters themselves, although we will not deal with this additional level of complexity in the current chapter.

Instead of using conjugate priors, one can opt for *noninformative* (or *vague*) priors in order to attenuate the impact on the resulting inference. These priors are indeed defined as extensions of the uniform distribution: For unbounded spaces, their densities actually fail to integrate to one and they are defined instead as positive measures. While this sounds like an invalid extension of the probabilistic framework, it is quite correct to define the corresponding posterior distributions by (2.2), as long as the integral in the denominator is finite (almost surely). The purpose of noninformative priors is in fact to provide a reference prior that has very little bearing on the inference (relative to the information brought by the likelihood). More detailed accounts are provided in Bernardo and Smith (1994) or Robert (2001, Section 1.5) about this possibility of using σ-finite measures (sometimes called

improper priors) in settings where true probability prior distributions are too difficult to come by or too subjective to be accepted by all. For instance, *location models*

$$y \sim p(y - \theta)$$

are usually associated with flat priors $\pi(\theta) = 1$, while *scale models*

$$y \sim \frac{1}{\theta} f\left(\frac{y}{\theta}\right)$$

are usually associated with the log-transform of a flat prior, that is,

$$\pi(\theta) = 1/\theta\,.$$

In a more general setting, the (noninformative) prior favored by most Bayesians is the so-called *Jeffreys' prior*,[3] which is related to Fisher's information $I^F(\theta)$ by

$$\pi^J(\theta) = \left|I^F(\theta)\right|^{1/2}\,,$$

where $|I|$ denotes the determinant of the matrix I.

Exercise 2.14. Show that, for location and scale models, Jeffreys' prior is given by $\pi^J(\theta) = 1$ and $\pi^J(\theta) = 1/\theta$, respectively.

Exercise 2.15. In the case of an exponential family, derive Jeffreys' prior in terms of the Hessian matrix of $\Psi(\theta)$, i.e., the matrix of second derivatives of $\Psi(\theta)$.

Since the mean μ of a normal model is a location parameter, the standard choice of noninformative parameter is $\pi(\mu) = 1$ (or any other constant). Given that this flat prior formally corresponds to the choice $\lambda = \mu = 0$ in the conjugate prior, it is easy to verify that this noninformative prior is associated with the posterior distribution $\mathcal{N}(x, 1)$, which happens to be the likelihood function in that case. An interesting consequence of this observation is that the MAP estimator is also the maximum likelihood estimator in that (special) case.

⊞ For **normaldata**, the posterior distributions associated with both the flat prior on μ and the conjugate prior $\mathcal{N}(0, 0.1\,\hat{\sigma}^2)$ are represented in Figure 2.4. The difference due to the two choices of priors is quite visible despite the 90 observations and illustrates the phenomenon known as the *lack of robustness* of the normal prior: When the hyperparameter (ξ, λ) varies, both the range and the location of the posterior distribution are not limited by the data.

[3] Harold Jeffreys was an English geophysicist who developed and formalized Bayesian methods in the 1930s in order to analyze geophysical data. He ended up writing an influential treatise on Bayesian statistics entitled *Theory of Probability*.

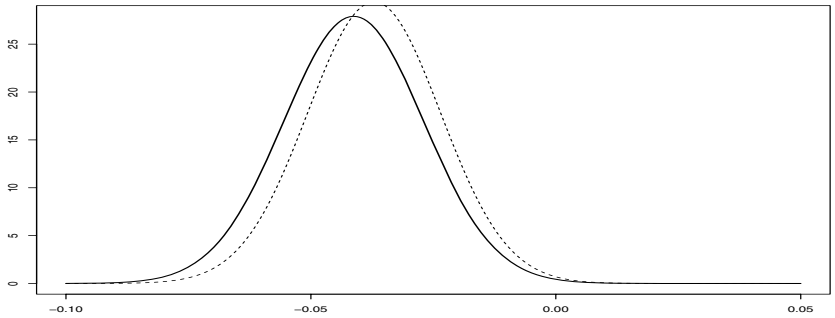


Fig. 2.4. Dataset normaldata: Two posterior distributions corresponding to the flat prior (plain) and a conjugate prior (dotted).

A major difference between proper and improper priors is that the posterior distribution associated with an improper prior is not necessarily defined, that is, it may happen that

$$\int \pi(\theta)\ell(\theta|\mathscr{D})\,\mathrm{d}\theta < \infty \tag{2.8}$$

does not hold. In some cases, this difficulty disappears when the sample size is large enough. In others (see Chapter 6), it may remain whatever the sample size. But the main thing is that, when using improper priors, condition (2.8) must always be checked.

Exercise 2.16. Show that, when $\pi(\theta)$ is a probability density, (2.8) necessarily holds for all datasets $\mathscr{D}$.

Exercise 2.17. Try to devise a parameterized model and an improper prior such that, no matter the sample size, the posterior distribution does not exist. (If you cannot find such a case, wait until Chapter 6.)

2.2.3 Confidence Intervals

One point that must be clear from the beginning is that the Bayesian approach is a complete inferential approach. Therefore, it covers confidence evaluation, testing, prediction, model checking, and point estimation. We will progressively cover the different facets of Bayesian analysis in other chapters of this book, but we address here the issue of confidence intervals.

As with everything else, the derivation of the confidence intervals (or confidence regions in more general settings) is based on the posterior distribution $\pi(\theta|\mathscr{D})$. Since the Bayesian approach processes θ as a random variable, a natural definition of a confidence region on θ is to determine $C(\mathscr{D})$ such that

$$\pi(\theta \in C(\mathscr{D})|\mathscr{D}) = 1 - \alpha \tag{2.9}$$

where α is a predetermined level such as 0.05.[4]

The important difference from a traditional perspective is that the integration is done over the parameter space, rather than over the observation space. The quantity $1-\alpha$ thus corresponds to the probability that a random θ belongs to this set $C(\mathscr{D})$, rather than to the probability that the random set contains the "true" value of θ. Given this drift in the interpretation of a confidence set (also called a *credible set* by Bayesians), the determination of the best[5] confidence turns out to be easier than in the classical sense: It simply corresponds to the values of θ with the highest posterior values,

$$C(\mathscr{D}) = \{\theta;\, \pi(\theta|\mathscr{D}) \geq k_\alpha\}\ ,$$

where k_α is determined by the coverage constraint (2.9). This region is called the *highest posterior density* (HPD) region.

⊞ For **normaldata**, using the parameter $\theta = (\mu, \sigma^2)$, the marginal posterior distribution on μ using the prior $\pi(\theta) = 1/\sigma^2$ is a Student's t distribution,

$$\pi(\mu|\mathscr{D}) \propto \left[n(\mu - \bar{x})^2 + s_x^2\right]^{-n/2}$$

with $n - 1 = 89$ degrees of freedom, as can be seen in (2.7), and the corresponding 95% confidence region for μ is the interval $[-0.070, -0.013]$. Note that, since 0 does not belong to this interval, one can feel justified in reporting a significant decrease in the number of larcenies between 1991 and 1995.

Ⓒ While the shape of an optimal Bayesian confidence set is easily derived, the computation of either the bound k_α or the set $C(\mathscr{D})$ may be too challenging to allow an analytic construction outside conjugate setups.

Example 2.1. When the prior distribution is not conjugate, the posterior distribution is not necessarily so well-behaved. For instance, if the normal $\mathscr{N}(\mu, 1)$ distribution is replaced with the Cauchy distribution, $\mathscr{C}(\mu, 1)$, in the likelihood

$$\ell(\mu|\mathscr{D}) = \prod_{i=1}^{n} f_\mu(x_i) = \frac{1}{\pi^n \prod_{i=1}^{n}(1 + (x_i - \mu)^2)}\ ,$$

there is no conjugate prior available and we can, for instance, consider a normal prior on μ, say $\mathscr{N}(0, 10)$. The posterior distribution is then proportional to

$$\tilde{\pi}(\mu|\mathscr{D}) = \frac{\exp(-\mu^2/20)}{\prod_{i=1}^{n}(1 + (x_i - \mu)^2)}\ .$$

[4] There is nothing special about 0.05 when compared with, say, 0.87 or 0.12. It is just that the famous 5% level is accepted by most as an acceptable level of error. If the loss function or the prior information tell a different story, another value for α (and one that may even depend on the data) should be chosen!

[5] In the sense of offering a given confidence coverage for the smallest possible length/volume.

Solving $\tilde{\pi}(\mu|\mathscr{D}) = k$ is not possible analytically, only numerically, and the derivation of the proper bound k_α does require an extra level of numerical computation in order to obtain the correct coverage. Figure 2.5 gives the posterior distribution of μ for the observations $x_1 = -4.3$ and $x_2 = 3.2$, normalized by simple trapezoidal integration, that is, by computing $\tilde{\pi}(\mu|\mathscr{D})$ on a regular grid of width Δ and summing up. For a given value of k, the same trapezoidal approximation can be used to compute the approximate coverage of the HPD region. For $\alpha = 0.95$, a trial-and-error exploration of the range of k then leads to an approximation of $k_\alpha = 0.0415$ and the HPD region, as represented in the figure. ◀

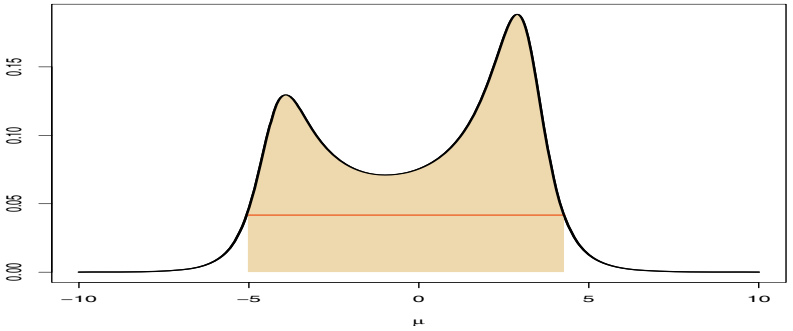


Fig. 2.5. Posterior distribution of μ for a $\mathscr{N}(0, 10)$ prior and 95% HPD region.

Given that posterior distributions are not necessarily unimodal, the HPD regions may include several disconnected sets, as illustrated by Example 2.1. This may seem counterintuitive from a classical point of view, but it must be interpreted as indicating indeterminacy, either in the data or in the prior, about the possible values of θ. Note also that HPDs are not independent from the choice of the reference measure that defines the volume (or surface).

2.3 Testing Hypotheses

Deciding the validity of some assumptions or restrictions on the parameter θ is a major part of the statistician's job. (We will deal with the assessment of the validity of the whole model—e.g., whether or not the normal distribution is appropriate for the data at hand—in Section 2.5.2 and more generally in Section 6.7.3.) Because the outcome of the decision process is clearcut, *accept* (coded by 1) or *reject* (coded by 0), the construction and the evaluation of procedures in this setup are quite crucial. While the Bayesian solution is formally very close from a likelihood ratio statistic, its numerical values often strongly differ from the classical solutions.

2.3.1 Zero–One Decisions

We will formalize assumptions as restricted parameter spaces, namely $\theta \in \Theta_0$. For instance, $\theta > 0$ corresponds to $\Theta_0 = \mathbb{R}^+$.

The standard Neyman–Pearson approach to hypothesis testing is based on the $0-1$ loss that equally penalizes all errors: If we consider the test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$, and denote by $d \in \{0, 1\}$ the decision made by the statistician and by δ the corresponding decision procedure, the loss

$$L(\theta, d) = \begin{cases} 1-d & \text{if} \quad \theta \in \Theta_0\,, \\ d & \text{otherwise,} \end{cases}$$

is associated with the risk

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[L(\theta, \delta(x))] \\ &= \int_{\mathcal{X}} L(\theta, \delta(x)) f_\theta(x) \mathrm{d}x \\ &= \begin{cases} P_\theta(\delta(x) = 0) & \text{if } \theta \in \Theta_0, \\ P_\theta(\delta(x) = 1) & \text{otherwise}\,, \end{cases} \end{aligned}$$

where x denotes the data $\mathscr{D}$ available. Under the $0-1$ loss, the Bayes decision (estimator) associated with a prior distribution π is

$$\delta^\pi(x) = \begin{cases} 1 & \text{if} \quad P^\pi(\theta \in \Theta_0 | x) > P^\pi(\theta \notin \Theta_0 | x), \\ 0 & \text{otherwise.} \end{cases}$$

This estimator is easily justified on an intuitive basis since it chooses the hypothesis with the largest posterior probability. A generalization of the loss above is to penalize errors differently when the null hypothesis is true than when it is false via the weighted $0-1$ losses $(a_0, a_1 > 0)$,

$$L_{a_0,a_1}(\theta, d) = \begin{cases} a_0 & \text{if} \quad \theta \in \Theta_0 \quad \text{and} \quad d = 0\,, \\ a_1 & \text{if} \quad \theta \in \Theta_1 \quad \text{and} \quad d = 1\,, \\ 0 & \text{otherwise.} \end{cases}$$

Exercise 2.18. Show that, under the loss L_{a_0,a_1}, the Bayes estimator associated with a prior π is given by

$$\delta^\pi(x) = \begin{cases} 1 & \text{if} \quad P^\pi(\theta \in \Theta_0 | x) > a_1 \big/ a_0 + a_1, \\ 0 & \text{otherwise.} \end{cases}$$

For this class of losses, the null hypothesis H_0 is rejected when the posterior probability of H_0 is too small, the acceptance level $a_1/(a_0 + a_1)$ being determined by the choice of (a_0, a_1).

Exercise 2.19. When $\theta \in \{\theta_0, \theta_1\}$, show that the Bayesian procedure only depends on the ratio $\varrho_0 f_{\theta_0}(x)/(1-\varrho_0) f_{\theta_1}(x)$, where ϱ_0 is the prior weight on θ_0.

⊞ For $x \sim \mathcal{N}(\mu, \sigma^2)$ and $\mu \sim \mathcal{N}(\xi, \tau^2)$, we recall that $\pi(\mu|x)$ is the normal distribution $\mathcal{N}(\xi(x), \omega^2)$ with

$$\xi(x) = \frac{\sigma^2\xi + \tau^2 x}{\sigma^2 + \tau^2} \quad \text{and} \quad \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

To test $H_0 : \mu < 0$, we compute

$$\begin{aligned} P^\pi(\mu < 0|x) &= P^\pi\left(\frac{\mu - \xi(x)}{\omega} < \frac{-\xi(x)}{\omega}\right) \\ &= \Phi\left(-\xi(x)/\omega\right). \end{aligned}$$

If z_{a_0,a_1} is the $a_1/(a_0+a_1)$ quantile of the normal $\mathcal{N}(0,1)$ distribution (i.e., $\Phi(z_{a_0,a_1}) = a_1/(a_0+a_1)$), H_0 is accepted when

$$-\xi(x) > z_{a_0,a_1}\omega,$$

the upper acceptance bound then being

$$x \leq -\frac{\sigma^2}{\tau^2}\xi - \left(1 + \frac{\sigma^2}{\tau^2}\right)\omega z_{a_0,a_1}.$$

This illustrates once more the *lack of robustness* of the conjugate prior: This bound can be any real value when ξ varies in $\mathbb{R}$.

Exercise 2.20. Show that the limit of the posterior probability $P^\pi(\mu < 0|x)$ when ξ goes to 0 and τ goes to ∞ is $\Phi(-x/\sigma)$.

2.3.2 The Bayes Factor

A notion central to Bayesian testing is the *Bayes factor*

$$B^\pi_{10} = \{P^\pi(\theta \in \Theta_1|x)/P^\pi(\theta \in \Theta_0|x)\} \,/\, \{P^\pi(\theta \in \Theta_1)/P^\pi(\theta \in \Theta_0)\}\,,$$

which corresponds to the classical odds or likelihood ratio, the difference being that the parameters are integrated rather than maximized under each model. While it is a simple one-to-one transform of the posterior probability, it can be used for Bayesian testing without resorting to a specific loss, for instance by using Jeffreys' scale of evidence:

- if $\log_{10}(B^\pi_{10})$ is between 0 and 0.5, the evidence against H_0 is *weak*,
- if it is between 0.5 and 1, it is *substantial*,

– if it is between 1 and 2, it is *strong*, and
– if it is above 2, it is *decisive*.

While this scale is far from justified on strict principles, it provides a reference for hypothesis assessment with no need to define the prior probabilities of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, which is one of the advantages of using the Bayes factor.

In general, the Bayes factor obviously depends on prior information, but it is still proposed as an objective Bayesian answer since it partly eliminates the influence of the prior modeling and emphasizes the role of the observations. Alternatively, it can be perceived as a Bayesian likelihood ratio since, if π_0 and π_1 are the prior distributions under H_0 and H_1, respectively, and if $P^\pi(\theta \in \Theta_0) = P^\pi(\theta \in \Theta_1) = 0.5$, B^π_{10} can be written as

$$B^\pi_{10} = \frac{\int_{\Theta_1} f_\theta(x)\pi_1(\theta)\,\mathrm{d}\theta}{\int_{\Theta_0} f_\theta(x)\pi_0(\theta)\,\mathrm{d}\theta} = \frac{m_1(x)}{m_0(x)}\,,$$

thus replacing the likelihoods with the marginals under both hypotheses.

When the hypothesis to be tested is a point null hypothesis, $H_0 : \theta = \theta_0$, there are difficulties in the construction of the Bayesian procedure, given that, for an absolutely continuous prior π,

$$P^\pi(\theta = \theta_0) = 0\,.$$

Obviously, point null hypotheses can be criticized as being artificial and impossible to test (*how often can one distinguish* $\theta = 0$ *from* $\theta = 0.0001$*?!*), but they must also be processed, being part of the everyday requirements of statistical analysis and also a convenient representation of some model choice problems (which we will discuss later).

Testing point null hypotheses actually requires a modification of the prior distribution so that, when testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$,

$$\pi(\Theta_0) > 0 \quad \text{and} \quad \pi(\Theta_1) > 0$$

hold, whatever the measures of Θ_0 and Θ_1 for the original prior, which means that the prior must be decomposed as

$$\pi(\theta) = P^\pi(\theta \in \Theta_0) \times \pi_0(\theta) + P^\pi(\theta \in \Theta_1) \times \pi_1(\theta)$$

with positive weights on both Θ_0 and Θ_1.

Note that this modification makes sense from both informational and operational points of view. If $H_0 : \theta = \theta_0$, the fact that the hypothesis is tested implies that $\theta = \theta_0$ *is* a possibility and it brings some additional prior information on the parameter θ. Besides, if H_0 is tested *and* accepted, this means that, in most situations, the (reduced) model under H_0 will be used rather than the (full) model considered before. Thus, a prior distribution under the

reduced model must be available for potential later inference. (Obviously, the fact that this later inference depends on the selection of H_0 should also be taken into account.)

In the special case $\Theta_0 = \{\theta_0\}$, π_0 is the Dirac mass at θ_0, which simply means that $P^{\pi_0}(\theta = \theta_0) = 1$, and we need to introduce a separate prior weight of H_0, namely,

$$\rho = P(\theta = \theta_0) \quad \text{and} \quad \pi(\theta) = \rho \mathbb{I}_{\theta_0}(\theta) + (1-\rho)\pi_1(\theta)\,.$$

Then,

$$\begin{aligned}\pi(\Theta_0|x) &= \frac{f_{\theta_0}(x)\rho}{\int f_\theta(x)\pi(\theta)\,\mathrm{d}\theta} \\ &= \frac{f_{\theta_0}(x)\rho}{f_{\theta_0}(x)\rho + (1-\rho)m_1(x)}.\end{aligned}$$

For $x \sim \mathcal{N}(\mu, \sigma^2)$ and $\mu \sim \mathcal{N}(\xi, \tau^2)$, consider the test of $H_0 : \mu = 0$. We can choose ξ equal to 0 if we do not have additional prior information. Then the Bayes factor is the ratio of marginals under both hypotheses, $\mu = 0$ and $\mu \neq 0$,

$$\begin{aligned}B_{10}^{\pi}(x) = \frac{m_1(x)}{f_0(x)} &= \frac{\sigma}{\sqrt{\sigma^2+\tau^2}} \frac{e^{-x^2/2(\sigma^2+\tau^2)}}{e^{-x^2/2\sigma^2}} \\ &= \sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}} \exp\left\{\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right\},\end{aligned}$$

and

$$\pi(\mu = 0|x) = \left[1 + \frac{1-\rho}{\rho}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}} \exp\left(\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right]^{-1}$$

is the posterior probability of H_0. Table 2.1 gives an indication of the values of the posterior probability when the normalized quantity x/σ varies. This posterior probability obviously depends on the choice of the prior variance τ^2: The dependence is actually quite severe, as we will see below with the *Jeffreys–Lindley paradox*.

Table 2.1. Posterior probability of $\mu = 0$ for different values of $z = x/\sigma$, $\rho = 1/2$, and for $\tau = \sigma$ (top), $\tau^2 = 10\sigma^2$ (bottom).

z	0	0.68	1.28	1.96
$\pi(\mu = 0\|z)$	0.586	0.557	0.484	0.351
$\pi(\mu = 0\|z)$	0.768	0.729	0.612	0.366

⊞ For **normaldata**, if we choose τ equal to 0.1, the Bayes factor against $\mu = 0$ only depends on $\overline{x}_n \sim \mathcal{N}(\mu, \hat{\sigma}^2/90)$ and is equal to

$$B_{10}^{\pi}(\overline{x}_n) = \sqrt{\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + n\tau^2}} \exp\left\{\frac{n\tau^2 \overline{x}_n^2}{2\hat{\sigma}^2(\hat{\sigma}^2 + n\tau^2)}\right\} = 0.1481\,.$$

The dataset is thus rather favorable to the null hypothesis $H_0 : \mu = 0$. However, other choices of τ lead to different numerical evaluations, as shown by Figure 2.6, since the Bayes factor varies between 1 and 0 as τ increases from 0 to ∞.

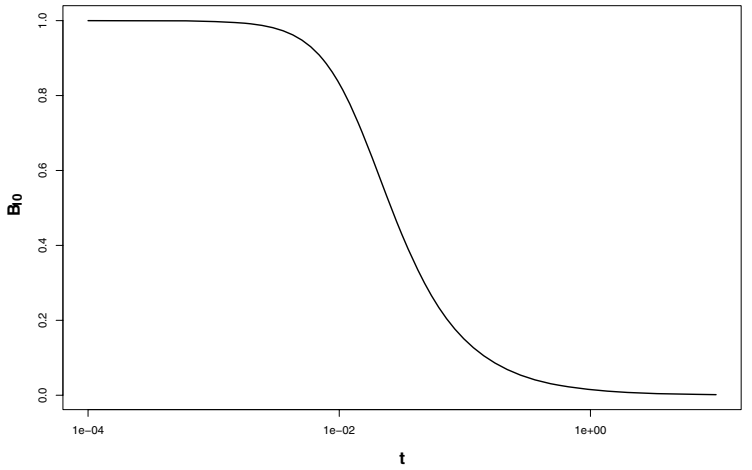


Fig. 2.6. Dataset normaldata: Range of the Bayes factor B_{10}^{π} when τ goes from 10^{-4} to 10. (*Note:* The x-axis is in logscale.)

2.3.3 The Ban on Improper Priors

Unfortunately, this decomposition of the prior distribution into two subpriors brings a serious difficulty related to improper priors, which amounts in practice to banning their use in testing situations. In fact, when using the representation

$$\pi(\theta) = P(\theta \in \Theta_0) \times \pi_0(\theta) + P(\theta \in \Theta_1) \times \pi_1(\theta)\,,$$

the weights $P(\theta \in \Theta_0)$ and $P(\theta \in \Theta_1)$ are meaningful only if π_0 and π_1 are normalized probability densities. Otherwise, they cannot be interpreted as *weights*.

When $x \sim \mathcal{N}(\mu, 1)$ and $H_0 : \mu = 0$, the improper (Jeffreys) prior is $\pi_1(\mu) = 1$; if we write

$$\pi(\mu) = \frac{1}{2}\mathbb{I}_0(\mu) + \frac{1}{2} \cdot 1\,,$$

then the posterior probability is

$$\pi(\mu = 0|x) = \frac{e^{-x^2/2}}{e^{-x^2/2} + \int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2}\,\mathrm{d}\theta} = \frac{1}{1 + \sqrt{2\pi}e^{x^2/2}}\,.$$

A first consequence of this choice is that the posterior probability of H_0 is bounded from above by

$$\pi(\mu = 0|x) \leq 1/(1 + \sqrt{2\pi}) = 0.285\,.$$

Table 2.2 provides the evolution of this probability as x goes away from 0. An interesting point is that the numerical values somehow coincide with the p-values used in classical testing (Casella and Berger, 2001).

Table 2.2. Posterior probability of $H_0 : \mu = 0$ for the Jeffreys prior $\pi_1(\mu) = 1$ under H_1.

x	0.0	1.0	1.65	1.96	2.58
$\pi(\mu = 0\|x)$	0.285	0.195	0.089	0.055	0.014

If we are instead testing $H_0 :\ \theta \leq 0$ versus $H_1 : \theta > 0$, then the posterior probability is

$$\pi(\theta \leq 0|x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-(x-\theta)^2/2}\,\mathrm{d}\theta = \Phi(-x)\,,$$

and the answer is now *exactly* the p-value found in classical statistics.

⊞ In the case of **normaldata**, if we consider the parameter $\theta = (\mu, \sigma^2)$ to be distributed from a noninformative $\pi(\theta) = 1/\sigma^2$ prior, the posterior probability that μ is positive is given as the probability that a Student's t distribution with 89 degrees of freedom and mean and variance -0.0144 and 0.000206, respectively, is positive. This is essentially a normal $\mathcal{N}(-0.0144, 0.000206)$ distribution, and the corresponding probability is 0.0021, which is very small for the hypothesis $H_0 : \mu > 0$ to hold in any likelihood.

The difficulty in using an improper prior also relates to what is called the *Jeffreys–Lindley paradox*, a phenomenon that shows that limiting arguments are not valid in testing settings. In contrast with estimation settings, the noninformative prior no longer corresponds to the limit of conjugate inferences. In fact, for a conjugate prior, the posterior probability

$$\pi(\theta = 0|x) = \left\{1 + \frac{1-\rho_0}{\rho_0}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left[\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right]\right\}^{-1}$$

converges to 1 when τ goes to $+\infty$, for *every* value of x, as already illustrated by Figure 2.6. This noninformative procedure obviously differs from the noninformative answer $[1+\sqrt{2\pi}\exp(x^2/2)]^{-1}$ above.

The fundamental issue that bars us from using improper priors on one or both of the sets Θ_0 and Θ_1 is a normalizing difficulty: If g_0 and g_1 are measures (rather than probabilities) on the subspaces Θ_0 and Θ_1, the choice of the normalizing constants influences the Bayes factor. Indeed, when g_i is replaced by $c_i g_i$ $(i = 0, 1)$, where c_i is an arbitrary constant, the Bayes factor is multiplied by c_0/c_1. Thus, for instance, if the Jeffreys prior is uniform and $g_0 = c_0$, $g_1 = c_1$, the posterior probability

$$\begin{aligned}\pi(\theta \in \Theta_0|x) &= \frac{\rho_0 c_0 \int_{\Theta_0} f(x|\theta)\,\mathrm{d}\theta}{\rho_0 c_0 \int_{\Theta_0} f(x|\theta)\,\mathrm{d}\theta + (1-\rho_0)c_1 \int_{\Theta_1} f(x|\theta)\,\mathrm{d}\theta} \\ &= \frac{\rho_0 \int_{\Theta_0} f(x|\theta)\,\mathrm{d}\theta}{\rho_0 \int_{\Theta_0} f(x|\theta)\,\mathrm{d}\theta + (1-\rho_0)[c_1/c_0] \int_{\Theta_1} f(x|\theta)\,\mathrm{d}\theta}\end{aligned}$$

is completely determined by the choice of c_0/c_1. This implies, for instance, that the function $[1+\sqrt{2\pi}\exp(x^2/2)]^{-1}$ obtained earlier has no validity whatsoever.

Since improper priors are an essential part of the Bayesian approach, there have been many proposals to overcome this ban. Most use a device that transforms the prior into a proper probability distribution by using a portion of the data $\mathscr{D}$ and then use the other part of the data to run the test as in a standard situation. The variety of available solutions is due to the many possibilities of removing the dependence on the choice of the portion of the data used in the first step. The resulting procedures are called *pseudo-Bayes factors*, although some may actually correspond to true Bayes factors. See Robert (2001, Chapter 6) for more details.

⊞ For **CMBdata**, consider two subsamples built by taking two segments on the image, as represented in Figure 2.7. We can consider that these subsamples, $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, both come from normal distributions, $\mathscr{N}(\mu_x, \sigma^2)$ and $\mathscr{N}(\mu_y, \sigma^2)$. The question of interest is to decide whether or not both means are identical, $H_0 : \mu_x = \mu_y$. To take advantage of the structures of this model, we can assume that σ^2 is the same measurement error under both models and thus that the same prior $\pi_\sigma(\sigma^2)$ can be used for both models. This means that the Bayes factor

$$B_{10}^{\pi} = \frac{\int \ell(\mu_x, \mu_y, \sigma|\mathscr{D})\pi(\mu_x, \mu_y)\pi_\sigma(\sigma^2)\,\mathrm{d}\sigma^2\,\mathrm{d}\mu_x\,\mathrm{d}\mu_y}{\int \ell(\mu, \sigma|\mathscr{D})\pi_\mu(\mu)\pi_\sigma(\sigma^2)\,\mathrm{d}\sigma^2\,\mathrm{d}\mu}$$

does not depend on the normalizing constant used for $\pi_\sigma(\sigma^2)$ and thus that we can still use an improper prior such as $\pi_\sigma(\sigma^2) = 1/\sigma^2$ in that case. Furthermore, we can rewrite μ_x and μ_y as $\mu_x = \mu - \xi$ and $\mu_y = \mu + \xi$, respectively, and use a prior of the form $\pi(\mu, \xi) = \pi_\mu(\mu)\pi_\xi(\xi)$ on the new parameterization so that, again, the same prior π_μ can be used under both H_0 and its alternative. The same cancellation of the normalizing constant occurs for π_μ, and the choice of prior $\pi_\mu(\mu) = 1$ and $\xi \sim \mathscr{N}(0, 1)$ leads to

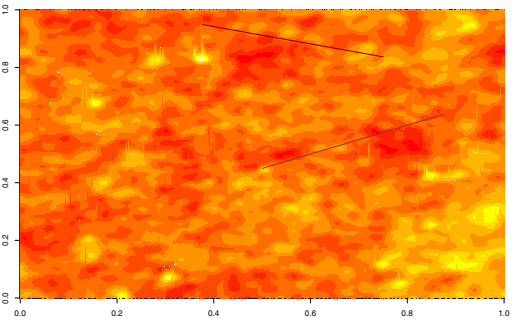


Fig. 2.7. Dataset **CMBdata**: Subsamples of the dataset represented by segments.

$$B_{10}^{\pi} = \frac{\int \exp\frac{-n}{2\sigma^2}\left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+S^2\right]\sigma^{-2n-2}e^{-\xi^2/2}/\sqrt{2\pi}\,\mathrm{d}\sigma^2\,\mathrm{d}\mu\,\mathrm{d}\xi}{\int \exp\frac{-n}{2\sigma^2}\left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+S^2\right]\sigma^{-2n-2}\,\mathrm{d}\sigma^2\,\mathrm{d}\mu}$$
$$= \frac{\int \left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+S^2\right]^{-n} e^{-\xi^2/2}/\sqrt{2\pi}\,\mathrm{d}\mu\,\mathrm{d}\xi}{\int \left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+S^2\right]^{-n}\mathrm{d}\mu}\,,$$

where S denotes the average

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\,.$$

While the denominator can be completely integrated out, the numerator cannot. A numerical approximation to B_{10}^{π} is thus necessary. This issue is addressed in Section 2.4.

Exercise 2.21. We recall that the normalizing constant for a Student's $\mathcal{T}(\nu,\mu,\sigma^2)$ distribution is

$$\frac{\Gamma((\nu+1)/2)/\Gamma(\nu/2)}{\sigma\sqrt{\nu\pi}}\,.$$

Give the value of the integral in the denominator of B_{10}^{π} above.

2.4 Monte Carlo Methods

While, as seen in Section 2.3, the Bayes factor and the posterior probability are the only quantities used in the assessment of hypotheses about the model, their analytical derivation is not always possible, since they involve integrating the parameter θ either on the set Θ_0 or on its complement Θ_0^c, under the respective

priors π_0 and π_1. There exist, however, special numerical techniques for the computation of Bayes factors, which are, mathematically speaking, simply ratios of integrals. We now detail the techniques used in the approximation of intractable integrals, but refer to Chen et al. (2000) and Robert and Casella (2004) for book-length presentations.

The technique that is most commonly used for integral approximations in statistics is called the Monte Carlo method[6] and relies on computer simulations of random variables to produce an approximation of integrals that converges with the number of simulations. Its justification is thus the *law of large numbers*, that is, if $x_1, \ldots, x_n$ are iid distributed from g, then the empirical average $\widehat{\mathfrak{I}}_n = (h(x_1) + \ldots + h(x_n))/n$ converges (almost surely) to the integral

$$\mathfrak{I} = \int h(x)g(x)\,\mathrm{d}x\,.$$

We will not expand on the foundations of the random number generators in this book, except for an introduction to accept–reject methods in Chapter 5, because of their links with Markov chain Monte Carlo techniques (see, instead, Robert and Casella, 2004). The connection of utmost relevance here is that these software packages necessarily have a limited collection of distributions available and that other methods must be found for simulating distributions outside this collection, paradoxically relying on the distributions already available, first and foremost the uniform $\mathscr{U}(0,1)$ distribution.

Ⓒ The implementation of the Monte Carlo method is straightforward, at least on a formal basis, with the following algorithmic representation:

Algorithm 2.1. Basic Monte Carlo Method

For $i = 1, \ldots, n$,

simulate $x_i \sim g(x)$.

Take

$$\widehat{\mathfrak{I}}_n = (h(x_1) + \ldots + h(x_n))/n$$

to approximate $\mathfrak{I}$.

as long as the (computer-generated) pseudo-random generation from g is feasible and the values $h(x_i)$ are computable. When simulation from g is a problem because g is nonstandard and usual techniques such as accept–reject algorithms (see Chapter 5) are difficult to devise, more advanced techniques such as Markov Chain Monte Carlo (MCMC) are required. We will introduce those in the next chapter. When the difficulty is with the intractability of the function h, the solution is often to use an integral representation of h and to expand the random variables x_i in (x_i, y_i), where y_i is an auxiliary variable. The use of such representations will be detailed in Chapter 6.

[6] This method is named in reference to the central district of Monaco, where the famous Monte-Carlo casino lies.

As computed in Exercise 2.21, the Bayes factor $B_{01}^\pi = 1/B_{10}^\pi$ can be simplified to give

$$B_{01}^\pi = \frac{\int \left[(\mu-\bar{x})^2 + (\mu-\bar{y})^2 + S^2\right]^{-n} \mathrm{d}\mu}{\int \left[(\mu-\xi-\bar{x})^2 + (\mu+\xi-\bar{y})^2 + S^2\right]^{-n} \, e^{-\xi^2/2}\,\mathrm{d}\mu\,\mathrm{d}\xi/\sqrt{2\pi}}$$
$$= \frac{\left[(\bar{x}-\bar{y})^2 + 2\,S^2\right]^{-n+1/2}}{\int \left[(2\xi+\bar{x}-\bar{y})^2 + 2\,S^2\right]^{-n+1/2} \, e^{-\xi^2/2}\,\mathrm{d}\xi/\sqrt{2\pi}}\,,$$

and we are left with a single integral in the denominator that involves the normal $\mathcal{N}(0,1)$ density and can thus be represented as an integral.

⊞ For both subsamples selected in Figure 2.7, we have

$$\bar{x} = 0.0888\,, \quad \bar{y} = 0.1078\,, \quad S^2 = 0.00875\,.$$

We thus simulate a $\mathcal{N}(0,1)$ sample of ξ_i's ($i = 1, \ldots, 1000$) and approximate B_{01}^π with

$$\widehat{B_{01}^\pi} = \frac{\left[(\bar{x}-\bar{y})^2 + 2\,S^2\right]^{-n+1/2}}{\frac{1}{1000}\sum_{i=1}^{1000}\left[(2\xi_i+\bar{x}-\bar{y})^2 + 2\,S^2\right]^{-n+1/2}}\,,$$

equal to 24.6 for the simulation experiment we ran. This value of $\widehat{B_{01}^\pi}$ means that H_0 is much more likely for the data at hand than $H_1 : \mu_x \neq \mu_y$. Note that, if we use a $\mathcal{N}(0, 0.01)$ prior instead, the approximated[7] value is 1.766, which considerably reduces the argument in favor of H_0.

Obviously, this *Monte Carlo estimate* of $\mathfrak{I}$ is not exact, but generating a sufficiently large number of random variables can render this approximation error arbitrarily small in a suitable probabilistic sense. It is even possible to assess the size of this error. If

$$\int |h(x)|^2 g(x)\,\mathrm{d}x < \infty\,,$$

the central limit theorem shows that $\sqrt{n}\,[\widehat{\mathfrak{I}}_n - \mathfrak{I}]$ is also normally distributed, and this can be used to construct asymptotic confidence regions for $\widehat{\mathfrak{I}}_n$, estimating the asymptotic variance from the simulation output.

⊞ For the approximation of B_{01}^π, the variation is illustrated in Figure 2.8, based on 1000 replications of the simulation of 1000 normal variables used in the approximation. As can be seen, the value of 24.6 reported in the previous analysis is rather central to the range of possible values, but the observed standard deviation of 19.5 shows that fairly different values could have been obtained as well (although all favor H_0, the minimum being 9.13). This shape of the histogram also shows that the normality promised by the central limit theorem is not yet attained.

[7] A change in the prior does not require a new simulation experiment since only the function h changes.

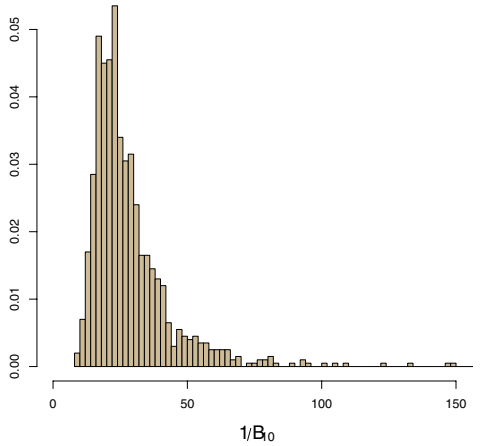


Fig. 2.8. Dataset CMBdata: Histogram of 1000 realizations of the approximation $\widehat{B_{10}}$ based on 1000 simulations each.

Exercise 2.22. Approximate B_{01}^{π} by a Monte Carlo experiment where ξ is simulated from a Student's t distribution with mean $(\bar{x}+\bar{y})/2$ and appropriate variance, and the integrand is proportional to $\exp -\xi^2/2$. Compare the precision of the resulting estimator with the Monte Carlo approximation above based on the normal simulation.

An important feature illustrated by Exercise 2.22 is that, for the Monte Carlo approximation of $\mathfrak{I}$, there is no need to simulate directly from g. In fact, $\mathfrak{I}$ can be represented in infinitely many ways as

$$\mathfrak{I} = \int \frac{h(x)g(x)}{\gamma(x)}\,\gamma(x)\,\mathrm{d}x\,, \tag{2.10}$$

where γ is another probability density. Therefore, the generation of a sample from γ can also provide an approximation to $\mathfrak{I}$ if the function $h(x)g(x)/\gamma(x)$ is used instead. This method is called *importance sampling* and applies in very wide generality.

While the representation (2.10) is true for any density γ with a support larger than the support of g, the performance of the empirical average $\widehat{\mathfrak{I}}_n$ may deteriorate considerably when the ratio $|h(x)g(x)/\gamma(x)|$ is not bounded because this opens a possibility of infinite variance in the resulting estimator. When using importance sampling, one must always take heed of the potential for infinite variance.

An additional incentive in using importance sampling is that this method does not require the density g to be known completely. It can simply be known up to a normalizing constant, $g(x) \propto \tilde{g}(x)$, since the ratio

$$\frac{\sum_{i=1}^{n} h(x_i)\tilde{g}(x_i)/\gamma(x_i)}{\sum_{i=1}^{n} \tilde{g}(x_i)/\gamma(x_i)}$$

also converges to $\mathfrak{I}$ when n goes to infinity and when the x_i's are generated from γ.

© The equivalent of Algorithm 2.1 for importance sampling is as follows:

ALGORITHM 2.2. IMPORTANCE SAMPLING METHOD
For $i = 1, \ldots, n$,
simulate $x_i \sim \gamma(x)$;
compute $\omega_i = \tilde{g}(x_i)/\gamma(x_i)$.
Take

$$\widehat{\mathfrak{I}}_n = \sum_{i=1}^{n} \omega_i \, h(x_i) \Big/ \sum_{i=1}^{n} \omega_i$$

to approximate $\mathfrak{I}$.

Once again, this algorithm is only formally straightforward to implement. But, since there is an additional degree of freedom in the selection of γ, simulation from this distribution can be imposed. And, in some settings, it can also be chosen so that the ratios $g(x_i)/\gamma(x_i)$ are easily computable.

Example 2.2. Consider almost the same setting as in Example 2.1, when $\mathscr{D} = (x_1, \ldots, x_n)$ is an iid sample from $\mathscr{C}(\theta, 1)$ and the prior on θ is now a flat prior. We can still use a normal importance function from the $\mathscr{N}(\mu, \sigma^2)$ distribution to produce a sample $\theta_1, \ldots, \theta_M$ that approximates the Bayes estimator of θ by

$$\hat{\delta}^{\pi}(\mathscr{D}) = \frac{\sum_{t=1}^{M} \theta_t \exp\left\{(\theta_t - \mu)^2/2\right\} \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}}{\sum_{t=1}^{M} \exp\left\{(\theta_t - \mu)^2/2\right\} \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}} \,.$$

But this is a very poor estimation (Exercise 2.24) and it degrades considerably when μ increases: As shown by Figure 2.9, not only does the range of the approximation increase, but it ends up missing the true value when μ is far enough from 0. ◀

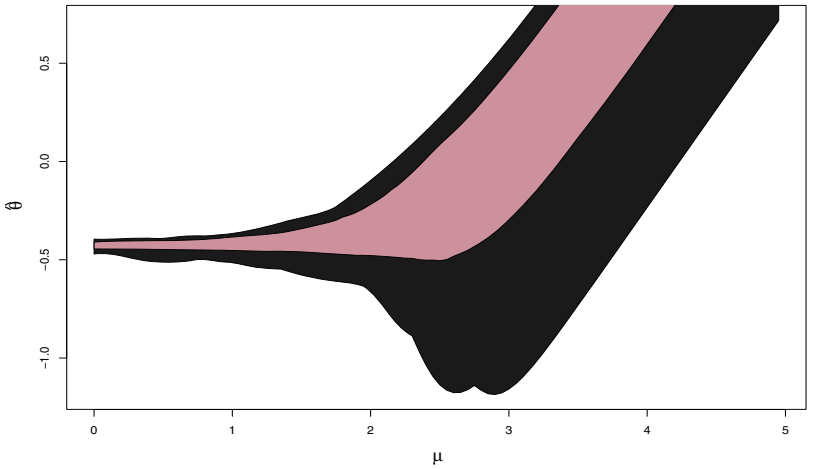


Fig. 2.9. Representation of the whole range (dark) and of the 90% range (grey) of variation of the importance sampling approximation to the Bayes estimate for $n = 10$ observations from the $\mathscr{C}(0,1)$ distribution and $M = 1000$ simulations of θ from a $\mathscr{N}(\mu,1)$ distribution as a function of μ. This range is computed using 1000 replications of the importance sampling estimates.

Exercise 2.23. Discuss what happens to the importance sampling approximation when the support of g is larger than the support of γ.

Exercise 2.24. Show that the importance weights of Example 2.2 have infinite variance.

Example 2.3. The density of Student's t distribution $\mathscr{T}(\nu,\theta,\sigma^2)$ is

$$f_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\frac{(x-\theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}.$$

(Without loss of generality, we will take $\theta = 0$, $\sigma = 1$.) Suppose that the integral of interest is

$$\mathfrak{I} = \int \sqrt{\left|\frac{x}{1-x}\right|}\, f_\nu(x)\,\mathrm{d}x\,,$$

which does exist. Then some possible choices for the importance function γ are

- the density of the Student $\mathscr{T}(\nu,0,1)$ distribution itself,
- the density of the Cauchy $\mathscr{C}(0,1)$ distribution, chosen for its heavy tails, and

– the density of the normal $\mathcal{N}(0, \nu/(\nu-2))$ distribution, chosen as a counter-example with light tails.

Exercise 2.25. Show that, when γ is the normal $\mathcal{N}(0, \nu/(\nu-2))$ density, the ratio

$$\frac{f_\nu^2(x)}{\gamma(x)} \propto \frac{e^{x^2(\nu-2)/2\nu}}{[1+x^2/\nu]^{(\nu+1)}}$$

does not have a finite integral. What does this imply about the variance of the importance weights?

The surprising feature with these different choices is that they all lead to very variable (and completely unreliable) estimates of $\mathfrak{I}$, as illustrated by Figure 2.10. The reason for this phenomenon is that the function h has a singularity at $x = 1$ such that

$$\int \frac{|x|}{|1-x|} f_\nu(x)\, \mathrm{d}x = \infty\,.$$

Therefore, the three estimators have infinite variance and we need to choose a better-behaved γ. Our next choice is then a modified gamma distribution, folded at 1, that is, the distribution of x is symmetric around 1 and is such that

$$|x-1| \sim \mathcal{G}a(\alpha, 1)\,.$$

Then

$$h(x)\frac{f^2(x)}{\gamma(x)} \propto \sqrt{x}\, f_\nu^2(x)\, |1-x|^{1-\alpha-1} \exp|1-x|$$

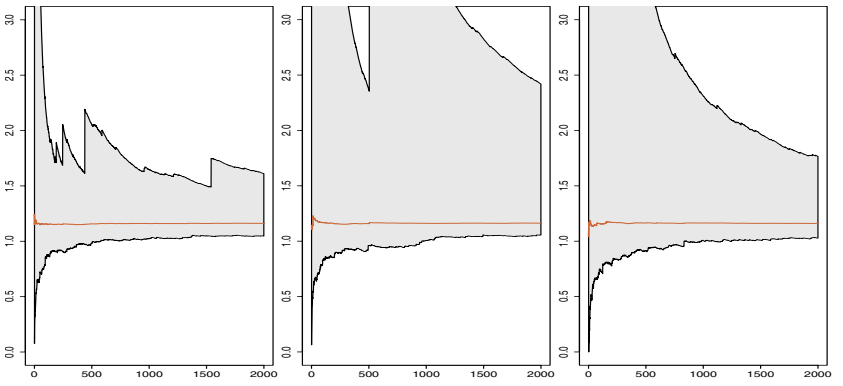


Fig. 2.10. Range of estimators $\hat{\mathfrak{I}}_n$ of $\mathfrak{I}$, as n increases from 0 to 2000, based on 500 replications of the sequences of $(\hat{\mathfrak{I}}_n)$: sampling from f_ν (left), a Cauchy distribution (center) and a normal distribution (right).

is integrable around $x = 1$ when $\alpha < 1$. (There is a problem of integrability at infinity, but this does not appear in the simulation experiments reproduced in Figure 2.11: The range over repeated simulations is well-behaved and shows the expected squared root decrease predicted by the central limit theorem.) ◀

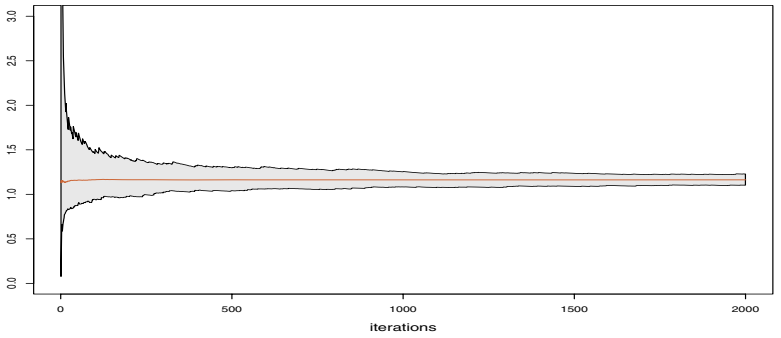


Fig. 2.11. Range of estimators $\hat{\mathfrak{I}}_n$ of $\mathfrak{I}$ based on 500 replications when sampling from the double gamma $\mathcal{G}a(1/2, 1)$ distribution folded at 1. (This graph is to be compared with Figure 2.10.)

Exercise 2.26. Given two model densities $f_1(\mathscr{D}|\theta)$ and $f_2(\mathscr{D}|\theta)$ with the same parameter θ and corresponding priors densities $\pi_1(\theta)$ and $\pi_2(\theta)$, denote $\tilde{\pi}_1(\theta|\mathscr{D}) = f_1(\mathscr{D}|\theta)\pi_1(\theta)$ and $\tilde{\pi}_2(\theta|\mathscr{D}) = f_2(\mathscr{D}|\theta)\pi_2(\theta)$, and show that the Bayes factor corresponding to the comparison of both models satisfies

$$B_{12}^{\pi} = \frac{\displaystyle\int \tilde{\pi}_1(\theta|\mathscr{D})\alpha(\theta)\pi_2(\theta|\mathscr{D})\mathrm{d}\theta}{\displaystyle\int \tilde{\pi}_2(\theta|\mathscr{D})\alpha(\theta)\pi_1(\theta|\mathscr{D})\mathrm{d}\theta}$$

for every positive function α and deduce that

$$n_1 \sum_{i=1}^{n_2} \tilde{\pi}_1(\theta_{2i}|\mathscr{D})\alpha(\theta_{2i}) \Big/ n_2 \sum_{i=1}^{n_1} \tilde{\pi}_2(\theta_{1i}|\mathscr{D})\alpha(\theta_{1i})$$

is a convergent approximation of the Bayes factor B_{12}^{π} when $\theta_{ji} \sim \pi_j(\theta|\mathscr{D})$ $(i = 1, 2,\ j = 1, \ldots, n_j)$.

2.5 Normal Extensions

The description of inference in normal models above is only an introduction both to Bayesian inference and to normal structures. Needless to say, there exists a much wider range of possible applications. For instance, we will meet the normal model again in Chapter 4 as the original case of the (generalized) linear model. Before that, we conclude this chapter with two simple extensions of interest: prediction inference and detection of outliers.

2.5.1 Prediction

When considering a sample $\mathscr{D}_n = (x_1, \ldots, x_n)$ from a normal $\mathscr{N}(\mu, \sigma^2)$ distribution, there can be a sequential or dynamic structure in the model that implies that future observations are expected. While more realistic modeling may involve probabilistic dependence between the x_i's, as presented in Chapter 7, we consider here the simpler setup of predictive distributions in iid settings.

If x_{n+1} is a future observation from the same distribution $f(\cdot|\theta)$ as the sample $\mathscr{D}_n$, its *predictive distribution* given the current sample is defined as

$$f^\pi(x_{n+1}|\mathscr{D}_n) = \int f(x_{n+1}|\theta, \mathscr{D}_n)\pi(\theta|\mathscr{D}_n)\,\mathrm{d}\theta = \int f(x_{n+1}|\theta)\pi(\theta|\mathscr{D}_n)\,\mathrm{d}\theta\,.$$

The motivation for defining this distribution is that the information available on the pair (x_{n+1}, θ) given the data $\mathscr{D}_n$ is summarized in the joint posterior distribution $f(x_{n+1}|\theta)\pi(\theta|\mathscr{D}_n)$ and the predictive distribution above is simply the corresponding marginal on x_{n+1}. This is obviously coherent with the Bayesian approach, which then considers x_{n+1} as an extra unknown.

Exercise 2.27. Show that, when n goes to infinity and when the prior has an unlimited support, the predictive distribution converges to the exact (sampling) distribution of x_{n+1}.

For the normal $\mathscr{N}(\mu, \sigma^2)$ setup, using a conjugate prior on (μ, σ^2) of the form

$$(\sigma^2)^{-\lambda_\sigma - 3/2}\,\exp - \left\{\lambda_\mu(\mu - \xi)^2 + \alpha\right\}/2\sigma^2\,,$$

the corresponding posterior distribution on (μ, σ^2) given $\mathscr{D}_n$ is

$$\mathscr{N}\left(\frac{\lambda_\mu \xi + n\overline{x}_n}{\lambda_\mu + n}, \frac{\sigma^2}{\lambda_\mu + n}\right) \times \mathscr{IG}\left(\lambda_\sigma + n/2, \left[\alpha + s_x^2 + \frac{n\lambda_\mu}{\lambda_\mu + n}(\overline{x} - \xi)^2\right]/2\right)\,,$$

denoted by

$$\mathscr{N}\left(\xi(\mathscr{D}_n), \sigma^2/\lambda_\mu(\mathscr{D}_n)\right) \times \mathscr{IG}\left(\lambda_\sigma(\mathscr{D}_n), \alpha(\mathscr{D}_n)/2\right)\,,$$

and the predictive on x_{n+1} is derived as

$$\begin{aligned} f^\pi(x_{n+1}|\mathscr{D}_n) &\propto \int (\sigma^2)^{-\lambda_\sigma-2-n/2} \exp -(x_{n+1}-\mu)^2/2\sigma^2 \\ &\qquad \times \exp -\left\{\lambda_\mu(\mathscr{D}_n)(\mu-\xi(\mathscr{D}_n))^2 + \alpha(\mathscr{D}_n)\right\}/2\sigma^2 \,\mathrm{d}(\mu,\sigma^2) \\ &\propto \int (\sigma^2)^{-\lambda_\sigma-n/2-3/2} \exp -\left\{(\lambda_\mu(\mathscr{D}_n)+1)(x_{n+1}-\xi(\mathscr{D}_n))^2 \right. \\ &\qquad \left. /\lambda_\mu(\mathscr{D}_n) + \alpha(\mathscr{D}_n)\right\}/2\sigma^2 \,\mathrm{d}\sigma^2 \\ &\propto \left[\alpha(\mathscr{D}_n) + \frac{\lambda_\mu(\mathscr{D}_n)+1}{\lambda_\mu(\mathscr{D}_n)}(x_{n+1}-\xi(\mathscr{D}_n))^2\right]^{-(2\lambda_\sigma+n+1)/2} . \end{aligned}$$

Therefore, the predictive of x_{n+1} given the sample $\mathscr{D}_n$ is a Student's t distribution with mean $\xi(\mathscr{D}_n)$ and $2\lambda_\sigma + n$ degrees of freedom. In the special case of the noninformative prior, $\lambda_\mu = \lambda_\sigma = \alpha = 0$ and the predictive is

$$f^\pi(x_{n+1}|\mathscr{D}_n) \propto \left[s_x^2 + \frac{n}{n+1}(x_{n+1}-\overline{x}_n)^2\right]^{-(n+1)/2} .$$

⊞ For **normaldata**, the (noninformative) predictive distribution on a 91st county (which we assume is distributed like the other 90 counties used in the study) is a Student's t with degrees of freedom 90, mean -0.0413, and scale parameter 0.136. This predictive is represented in Figure 2.12, which shows that the range of the predictive agrees rather closely with the range of the dataset, except for one extreme value, already mentioned.

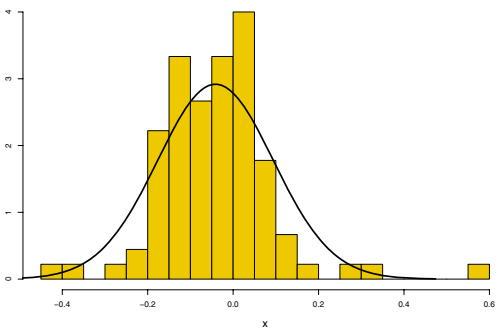


Fig. 2.12. Dataset normaldata: Predictive distribution compared with its histogram.

2.5.2 Outliers

Since normal modeling is often an approximation to the "real thing," there may be doubts about its adequacy. As already mentioned above, we will deal

later with the problem of checking that the normal distribution is appropriate for the whole dataset. Here, we consider the somehow simpler problem of assessing whether or not each point in the dataset is compatible with normality. There are many different ways of dealing with this problem. We choose here to take advantage of the derivation of the predictive distribution above: If an observation x_i is unlikely under the predictive distribution based on the *other observations*, then we can argue against its distribution being equal to the distribution of the other observations.

For each $x_i \in \mathcal{D}_n$, we consider $f_i^\pi(x|\mathcal{D}_n^i)$ as being the predictive distribution based on $\mathcal{D}_n^i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Considering $f_i^\pi(x_i|\mathcal{D}_n^i)$ or the corresponding cdf $F_i^\pi(x_i|\mathcal{D}_n^i)$ (in dimension one) gives an indication of the level of compatibility of the observation with the sample. To quantify this level, we can, for instance, approximate the distribution of $F_i^\pi(x_i|\mathcal{D}_n^i)$ as uniform over $[0,1]$ since (Exercise 2.27) $F_i^\pi(\cdot|\mathcal{D}_n^i)$ converges to the true cdf of the model. Simultaneously checking all $F_i^\pi(x_i|\mathcal{D}_n^i)$ over i may signal outliers.

Exercise 2.28. Show that, when X is distributed from an increasing and continuous cdf F, $F(X)$ has a uniform distribution.

↯ The detection of outliers must pay attention to the *Bonferroni fallacy*, which is that extreme values do occur in large enough samples. This means that, as n increases, we will see smaller and smaller values of $F_i^\pi(x_i|\mathcal{D}_n^i)$ even if the whole sample is from the same distribution. The significance level must therefore be chosen in accordance with this observation, for instance using a bound a on $F_i^\pi(x_i|\mathcal{D}_n^i)$ such that

$$1 - (1-a)^n = 1 - \alpha\,,$$

where α is the nominal level chosen for outlier detection.

⊞ Using **normaldata**, we can compute the predictive cdf for each of the 90 observations. Figure 2.13 provides the qq-plot of the $F_i^\pi(x_i|\mathcal{D}_n^i)$'s against the uniform quantiles and compares it with a qq-plot based on a dataset truly simulated from the uniform $\mathcal{U}(0,1)$: There is no clear departure from uniformity when looking at this graph.

When looking at the extreme values of $F_i^\pi(x_i|\mathcal{D}_n^i)$ (or at the smallest values of $f_i^\pi(x_i|\mathcal{D}_n^i)$), we can nonetheless detect local apparent departures. For instance, the smallest $F_i^\pi(x_i|\mathcal{D}_n^i)$ is equal to 0.00174: The probability that a uniform sample of 90 points has at least one value below 0.00174 is 0.146, and 0.00174 is thus compatible with the other points of the sample.

The largest $F_i^\pi(x_i|\mathcal{D}_n^i)$ is equal to 0.9999996: The probability that a uniform sample of 90 points has at least one value above this value is 0.0000334, and the corresponding x_i is thus unlikely to have the same distribution as the other counties. (The second-largest value is 0.996, which is completely acceptable for a uniform sample of 90 points.)

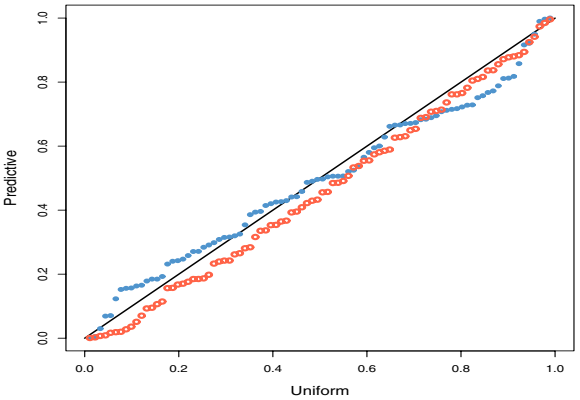


Fig. 2.13. Dataset normaldata: qq-plot of the sample of the $F_i^\pi(x_i|\mathcal{D}_n^i)$ for a uniform $\mathscr{U}(0,1)$ distribution (full dots) and comparison with a qq-plot for a uniform $\mathscr{U}(0,1)$ sample (empty dots).

3

Regression and Variable Selection

You see, I always keep my sums.
—**Ian Rankin**, ***Strip Jack***.—

Roadmap

Linear regression is one of the most widely used tools in statistics for analyzing the (linear) influence of some variables or some factors on others and thus to uncover explanatory and predictive patterns. This chapter unfolds the Bayesian analysis of the linear model both in terms of prior specification (conjugate, noninformative, and Zellner's G-prior) and in terms of variable selection, the next chapter appearing as a sequel for nonlinear dependence structures. The reader should be warned that, given that these models are the only conditional models where explicit computation can be conducted, this chapter contains a fair amount of matrix calculus. The photograph at the top of this page is a picture of processionary caterpillars, in connection (for once!) with the benchmark dataset used in this chapter. (This picture was taken with permission from the Website of Professor T.D. Fitzgerald, SUNY Cortland.)

3.1 Linear Dependence

A large proportion of statistical analyses deal with the representation of dependences among several observed quantities. For instance, which social factors influence unemployment duration and the probability of finding a new job? Which economic indicators are best related to recession occurrences? Which physiological levels are most strongly correlated with aneurysm strokes? From a statistical point of view, the ultimate goal of these analyses is thus to find a proper representation of the conditional distribution, $f(y|\theta, \mathbf{x})$, of an observable variable y given a vector of observables $\mathbf{x}$, based on a sample of $\mathbf{x}$ and y. While the overall estimation of the conditional density f is usually beyond our ability, the estimation of θ and possibly of restricted features of f is possible within the Bayesian framework, as shown in this chapter.

The variable of primary interest, y, is called the *response* or the *outcome* variable; we assume here that this variable is continuous, but we will completely relax this assumption in the next chapter. The variables $\mathbf{x} = (x_1, \ldots, x_k)$ are called *explanatory variables* and may be discrete, continuous, or both. One sometimes picks a single variable x_j to be of primary interest. We then call it the *treatment* variable, labeling the other components of x as *control* variables, meaning that we want to address the influence of x_j on y once the influence of all the other variables has been taken into account (as in medical studies). The distribution of y given $\mathbf{x}$ is typically studied in the context of a set of *units* or experimental *subjects*, $i = 1, \ldots, n$, such as patients in a hospital ward, on which both y_i and $x_{i1}, \ldots, x_{ik}$ are measured. The dataset is then made up of the conjunction of the vector of outcomes

$$\mathbf{y} = (y_1, \ldots, y_n)$$

and the $n \times (k+1)$ matrix of explanatory variables

$$X = [\mathbf{1}_n \quad \mathbf{x}_1 \quad \ldots \quad \mathbf{x}_k] = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ 1 & x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} .$$

(We will see very soon why it is better to add a column of 1's to the genuine explanatory variables.)

⊞ The **caterpillar** dataset used in this chapter was extracted from a 1973 study on pine processionary[1] caterpillars: It assesses the influence of some forest settlement characteristics on the development of caterpillar colonies. (It was published and studied in

[1] These caterpillars got their name from their habit of moving over the ground in incredibly long head-to-tail processions when leaving their nest to create a new colony.

Tomassone et al., 1993.) The response variable is the logarithmic transform of the average number of nests of caterpillars per tree in an area of 500 square meters (which corresponds to the last column in **caterpillar**). There are $k = 10$ potential explanatory variables defined on $n = 33$ areas, as follows

x_1 is the altitude (in meters),
x_2 is the slope (in degrees),
x_3 is the number of pines in the area,
x_4 is the height (in meters) of the tree sampled at the center of the area,
x_5 is the diameter of the tree sampled at the center of the area,
x_6 is the index of the settlement density,
x_7 is the orientation of the area (from 1 if southbound to 2 otherwise),
x_8 is the height (in meters) of the dominant tree,
x_9 is the number of vegetation strata,
x_{10} is the mix settlement index (from 1 if not mixed to 2 if mixed).

The goal of the regression analysis is to decide which explanatory variables have a strong influence on the number of nests and how these influences overlap with one another. As shown by Figure 3.1, some of these variables clearly have a restricting influence on the number of nests, as for instance with x_1, x_2, x_8, and x_9.

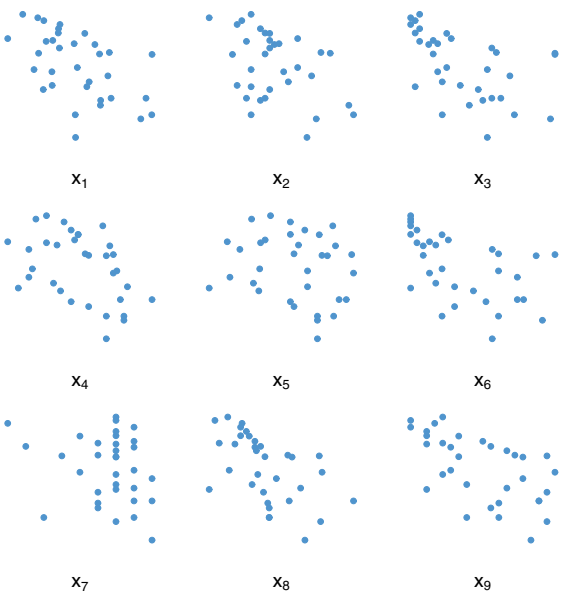


Fig. 3.1. Dataset caterpillar: Plot of the pairs $(\mathbf{x}_i, \mathbf{y})$ $(1 \leq i \leq 10)$.

3.1.1 Linear Models

While many models and thus many dependence structures can be proposed for dependent datasets like caterpillar, in this chapter we only focus on the Gaussian linear regression model, namely the case when $\mathbb{E}[y|x,\theta]$ is linear in x and the noise is normal.

The *ordinary normal linear regression* model is such that

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

and thus

$$\mathbb{E}[y_i|\beta,X] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\,, \quad \mathbb{V}(y_i|\sigma^2,X) = \sigma^2\,.$$

In particular, the presence of an *intercept* β_0 explains why a column of 1's is necessary in the matrix X to preserve the compact formula $X\beta$ in the conditional mean of $\mathbf{y}$.

⊞ For caterpillar, where $n = 33$ and $k = 10$, we thus assume that the expected lognumber y_i of caterpillar nests per tree over an area is modeled as a linear combination of an intercept and ten predictor variables ($i = 1,\ldots,n$),

$$\mathbb{E}[y_i|\beta,X] = \beta_0 + \sum_{j=1}^{10} \beta_j x_{ij}\,,$$

while the variation around this expectation is supposed to be normally distributed. Note that it is also customary to assume that the y_i's are conditionally independent.

There is a difference between using finite-valued regressors like x_7 in caterpillar and using *categorical* variables (or *factors*), which also take a finite number of values but whose range has no numerical meaning. For instance, if x denotes the socio-professional category of an employee, this variable may range from 1 to 9 for a rough grid of socio-professional activities, or it may range from 1 to 89 on a finer grid, and the numerical values are not comparable. It thus makes little sense to involve x directly in the regression, and the usual approach is to replace the single regressor x (taking values in $\{1,\ldots,m\}$, say) with m indicator (or *dummy*) variables $x_1 = \mathbb{I}_1(x)$, $\ldots$, $x_m = \mathbb{I}_m(x)$. In essence, a different constant (or *intercept*) β_i is used in the regression for each class of categorical variable: It is invoked in the linear regression under the form

$$\ldots + \beta_1 \mathbb{I}_1(x) + \ldots + \beta_m \mathbb{I}_m(x) + \ldots\,.$$

Obviously, there is an identifiability issue there since the sum of the indicators is equal to one. In a Bayesian approach, identifiability can be achieved via the prior distribution, but we can also impose an identifiability constraint on the parameters, for instance the omission of one class (such as $\beta_1 = 0$). We develop this remark further in Sections 4.5.1 and 6.2.

3.1.2 Classical Estimators

Before launching into the description of the Bayesian approach to the linear model, we recall the basics of the classical processing of this model (in particular, to relate the Bayesian perspective to the results provided by standard software such as R lm output). For instance, the parameter β can obviously be estimated via maximum likelihood estimation. In order to avoid non-identifiability and uniqueness problems, we assume that X is of full rank, that is, $\text{rank}(X) = k+1$. (This also means that there is no redundant structure among the explanatory variables.)[2] We suppose in addition that $k+1 < n$ in order to obtain proper estimates of all parameters.

Exercise 3.1. Show that the matrix X is of full rank if and only if the matrix $X^{\mathsf{T}}X$ is invertible (where X^{T} denotes the transpose of the matrix X, which can be produced in R using the `t(X)` command). Deduce that this cannot happen when $k+1 > n$.

The likelihood of the *ordinary normal linear model* is

$$\ell\left(\beta, \sigma^2 | \mathbf{y}, X\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - X\beta)^{\mathsf{T}}(\mathbf{y} - X\beta)\right]. \tag{3.1}$$

The maximum likelihood estimator of β is then the solution of the (least squares) minimization problem

$$\min_{\beta} (\mathbf{y} - X\beta)^{\mathsf{T}}(\mathbf{y} - X\beta) = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}\right)^2 ,$$

namely,

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y} ,$$

which is also the orthogonal projection of $\mathbf{y}$ on the linear subspace spanned by the columns of X (Exercise 3.2).

Exercise 3.2. Show that solving the minimization program above requires solving the system of equations $(X^{\mathsf{T}}X)\beta = X^{\mathsf{T}}\mathbf{y}$. Check that this can be done via the R command

```
> solve(t(X)%*%(X),t(X)%*%y)
```

It is quite simple to check that $\hat{\beta}$ is an unbiased estimator of β. Moreover, the Gauss–Markov theorem (Christensen, 2002) states that $\hat{\beta}$ is the *best* linear unbiased estimator of β. This means that, for all $a \in \mathbb{R}^{k+1}$, $\mathbb{V}(a^{\mathsf{T}}\hat{\beta}|\sigma^2, X) \leq \mathbb{V}(a^{\mathsf{T}}\tilde{\beta}|\sigma^2, X)$ for any unbiased linear estimator $\tilde{\beta}$ of β. (Of course, the property of unbiasedness per se is not particularly appealing.)

[2] Hence, the exclusion of one class for categorical variables.

Exercise 3.3. Show that $\mathbb{V}(\hat{\beta}|\sigma^2, X) = \sigma^2(X^\mathsf{T}X)^{-1}$.

Similarly, an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(\mathbf{y}-\hat{\beta})^\mathsf{T}(\mathbf{y}-\hat{\beta}) = \frac{s^2}{n-k-1},$$

and $\hat{\sigma}^2(X^\mathsf{T}X)^{-1}$ approximates the covariance matrix of $\hat{\beta}$. We can then define the standard *t-statistic* as

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2\omega_{(i,i)}}} \sim \mathscr{T}_{n-k-1}$$

where $\omega_{(i,i)}$ denotes the (i,i)th element of the matrix $(X^\mathsf{T}X)^{-1}$.

This t-statistic is then used in classical tests, for instance to accept $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ at the level α if $|\hat{\beta}_i|/\sqrt{\omega_{(i,i)}} < F^{-1}_{n-k-1}(1-\alpha/2)$, the $(1-\alpha/2)$nd quantile of the $\mathscr{T}_{n-k-1}$ distribution. The frequentist argument in using this bound is that the so-called *p-value* is smaller than α,

$$p_i = \mathbb{P}_{H_0}(|T_i| > |t_i|) < \alpha.$$

Note that this statistic T_i can also be used to build on the β_is a (marginal) frequentist confidence interval, of the form

$$\left\{\beta_i;\ \left|\beta_i - \hat{\beta}_i\right| \leq \sqrt{\omega_{ii}}\, F^{-1}_{n-k-1}(1-\alpha/2)\right\}.$$

Needless to say, we do not advocate the use of p-values in Bayesian settings or elsewhere since it involves many severe defects (exposed for instance in Robert, 2001, Chapter 5), one being that it is often wrongly interpreted as the probability of the null hypothesis.

For caterpillar, the unbiased estimate of σ^2 is equal to 0.688. The maximum likelihood estimates of the components β_i produced by the R command

```
> summary(lm(y~x))
```

are given in Figure 3.2, along with the estimates of their respective standard deviations and p-values. According to the classical paradigm, the coefficients β_3, $\beta_6, \ldots, \beta_{10}$ are therefore not *significant*.

For future use, we also note that the likelihood (3.1) satisfies

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.6989839 -0.2731726 -0.0003620  0.3246311  1.7304969

Coefficients:
            Estimate  Std. Error  t-value Pr(>|t|)
intercept   10.998412   3.060272   3.594  0.00161 **
XV1         -0.004431   0.001557  -2.846  0.00939 **
XV2         -0.053830   0.021900  -2.458  0.02232 *
XV3          0.067939   0.099472   0.683  0.50174
XV4         -1.293636   0.563811  -2.294  0.03168 *
XV5          0.231637   0.104378   2.219  0.03709 *
XV6         -0.356800   1.566464  -0.228  0.82193
XV7         -0.237469   1.006006  -0.236  0.81558
XV8          0.181060   0.236724   0.765  0.45248
XV9         -1.285316   0.864847  -1.486  0.15142
XV10        -0.433106   0.734869  -0.589  0.56162
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 3.2. Dataset caterpillar: R output providing the maximum likelihood estimates of the regression coefficients and their standard significance analysis.

$$\ell\left(\beta,\sigma^2|\mathbf{y},X\right)=\left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-X\hat{\beta})^{\mathsf{T}}(\mathbf{y}-X\hat{\beta})\right.$$
$$\left.-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^{\mathsf{T}}(X^{\mathsf{T}}X)(\beta-\hat{\beta})\right], \tag{3.2}$$

given that

$$\begin{aligned}(\mathbf{y}-X\hat{\beta})^{\mathsf{T}}X(\beta-\hat{\beta})&=\mathbf{y}^{\mathsf{T}}(I_n-X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}})X(\beta-\hat{\beta})\\&=\mathbf{y}^{\mathsf{T}}(X-X)(\beta-\hat{\beta})\\&=0\end{aligned}$$

by virtue of the projection argument above. Therefore, $s^2=(\mathbf{y}-X\hat{\beta})^{\mathsf{T}}(\mathbf{y}-X\hat{\beta})$ (or $\hat{\sigma}^2$) and $\hat{\beta}$ constitute a sufficient statistic for this model.

Conditioning on X is valid only if X is *exogenous*, that is, only if we can write the joint distribution of $(\mathbf{y},X)$ as

$$f(\mathbf{y},X|\beta,\sigma^2,\alpha)=f(\mathbf{y}|\beta,\sigma^2,X)f(X|\alpha)\,,$$

where (β,σ^2) and α are fixed parameters. We can thus ignore $f(X|\alpha)$ if the parameter α is only a nuisance parameter since this part is independent[3]of (β,σ^2). The practical advantage of using a regression model as above is that it is much easier to specify a realistic conditional distribution of one variable

given k others rather than a joint distribution on all $k+1$ variables. Note that if X is not *exogenous*, for instance when X involves past values of $\mathbf{y}$ (see Chapter 7), the joint distribution must be used instead.

3.2 First-Level Prior Analysis

As we now consider the Bayesian approach to this model, a careful analysis of the construction of the corresponding prior distributions is necessary. The current section deals with the conjugate first-level case, while Section 3.3 focuses on the extension to the noninformative case.[4]

3.2.1 Conjugate Priors

The likelihood shape (3.2) is Gaussian/Inverse Gamma: Specifically, β given σ^2 appears in a Gaussian-like expression, while σ^2 suggests an Inverse Gamma expression. The conjugate prior can thus be constructed within this family. Specifying both the conditional prior on β,

$$\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}),$$

where M is a $(k+1, k+1)$ positive definite symmetric matrix, and the marginal prior on σ^2,

$$\sigma^2|X \sim \mathcal{IG}(a, b), \qquad a, b > 0,$$

we indeed have conjugate priors in that the conditional posterior distribution $\pi(\beta|\sigma^2, \mathbf{y}, X)$ is

$$\mathcal{N}_{k+1}\left((M + X^{\mathsf{T}}X)^{-1}\{(X^{\mathsf{T}}X)\hat{\beta} + M\tilde{\beta}\}, \sigma^2(M + X^{\mathsf{T}}X)^{-1}\right), \tag{3.3}$$

and the marginal posterior distribution $\pi(\sigma^2|\mathbf{y}, X)$ is

$$\mathcal{IG}\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{(\tilde{\beta} - \hat{\beta})^{\mathsf{T}}\left(M^{-1} + (X^{\mathsf{T}}X)^{-1}\right)^{-1}(\tilde{\beta} - \hat{\beta})}{2}\right), \tag{3.4}$$

which are of the same types as the prior distributions.

[3] Obviously, from a Bayesian point of view, we would also have to impose prior independence between (β, σ^2) and α to achieve this separation.

[4] In order to keep this presentation simple and self-contained, we make several choices in the presentation that the most mature readers will possibly find arbitrary, but this cannot be avoided if we want to keep the chapter at a reasonable length.

Exercise 3.4. Taking advantage of the matrix identities

$$\begin{aligned}\left(M + X^{\mathsf{T}}X\right)^{-1} &= M^{-1} - M^{-1}\left(M^{-1} + (X^{\mathsf{T}}X)^{-1}\right)^{-1} M^{-1} \\ &= (X^{\mathsf{T}}X)^{-1} - (X^{\mathsf{T}}X)^{-1}\left(M^{-1} + (X^{\mathsf{T}}X)^{-1}\right)^{-1}(X^{\mathsf{T}}X)^{-1}\end{aligned}$$

and

$$\begin{aligned}X^{\mathsf{T}}X(M + X^{\mathsf{T}}X)^{-1}M &= \left(M^{-1}(M + X^{\mathsf{T}}X)(X^{\mathsf{T}}X)^{-1}\right)^{-1} \\ &= \left(M^{-1} + (X^{\mathsf{T}}X)^{-1}\right)^{-1},\end{aligned}$$

establish that (3.3) and (3.4) are the correct posterior distributions.

In this setting, the Bayes estimators of β and σ^2 associated with squared error losses, namely the posterior means of β and σ^2, can be computed in closed form. In fact, a simple computation shows that they are given by

$$\begin{aligned}\mathbb{E}[\beta|\mathbf{y}, X] &= \mathbb{E}\left[\mathbb{E}(\beta|\sigma^2, \mathbf{y}, X)|\mathbf{y}, X\right] \\ &= (M + X^{\mathsf{T}}X)^{-1}\{(X^{\mathsf{T}}X)\hat{\beta} + M\tilde{\beta}\} \qquad (3.5)\end{aligned}$$

and (for $n \geq 2$)

$$\mathbb{E}[\sigma^2|\mathbf{y}, X] = \frac{2b + s^2 + (\tilde{\beta} - \hat{\beta})^{\mathsf{T}}\left\{M^{-1} + (X^{\mathsf{T}}X)^{-1}\right\}^{-1}(\tilde{\beta} - \hat{\beta})}{n + 2a - 2}.$$

There is a close connection between the Bayes estimator (3.5) of β and *ridge regression*, which is sometimes used in standard regression packages to numerically stabilize unstable maximum likelihood estimates, that is, when the matrix $X^{\mathsf{T}}X$ is close to being noninvertible (this imprecisely defined situation is also called *near multicollinearity*). Indeed, the ridge estimator is defined as

$$(I_{k+1}/c + X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$$

and depends on the stabilization factor $c > 0$. Setting $M = I_{k+1}/c$ and $\tilde{\beta} = 0_{k+1}$ in the prior distribution, we see that this estimator is identical to the posterior mean. The general Bayes estimator (3.5) is thus a weighted average between the maximum likelihood estimator and the prior mean.

Integrating (3.3) over (3.4) in σ^2 leads to a multivariate t marginal posterior distribution on β since

$$\begin{aligned}\pi(\beta|\mathbf{y}, X) \propto \Big[&(\beta - \{M + X^{\mathsf{T}}X\}^{-1}[\{X^{\mathsf{T}}X\}\hat{\beta} + M\tilde{\beta}])^{\mathsf{T}}(M + X^{\mathsf{T}}X) \\ &\times(\beta - \{M + X^{\mathsf{T}}X\}^{-1}[\{X^{\mathsf{T}}X\}\hat{\beta} + M\tilde{\beta}]) + 2b + s^2 \\ &+ (\tilde{\beta} - \hat{\beta})^{\mathsf{T}}\left(M^{-1} + (X^{\mathsf{T}}X)^{-1}\right)^{-1}(\tilde{\beta} - \hat{\beta})\Big]^{-(n/2+k/2+a)}\end{aligned}$$

(the computation is straightforward if tedious bookkeeping). We recall that the density of a multivariate $\mathscr{T}_p(\nu, \theta, \Sigma)$ distribution is

$$f(t|\nu,\theta,\Sigma) = \frac{\Gamma((\nu+p)/2)/\Gamma(\nu/2)}{\sqrt{\det(\Sigma)\nu\pi}} \left[1 + \frac{(t-\theta)^\mathsf{T}\Sigma^{-1}(t-\theta)}{\nu}\right]^{-(\nu+p)/2}.$$

We thus have that, marginally and a posteriori,

$$\beta|\mathbf{y}, X \sim \mathscr{T}_{k+1}\left(n+2a, \hat{\mu}, \hat{\Sigma}\right),$$

with

$$\begin{aligned}\hat{\mu} &= (M + X^\mathsf{T}X)^{-1}((X^\mathsf{T}X)\hat{\beta} + M\tilde{\beta}), \qquad (3.6)\\ \hat{\Sigma} &= \frac{2b + s^2 + (\tilde{\beta}-\hat{\beta})^\mathsf{T}\left(M^{-1} + (X^\mathsf{T}X)^{-1}\right)^{-1}(\tilde{\beta}-\hat{\beta})}{n+2a}(M + X^\mathsf{T}X)^{-1}.\end{aligned}$$

In this case, the posterior variance of β is proportional to $\left(M^{-1} + X^\mathsf{T}X\right)^{-1}$. The scale factor comes from the Inverse Gamma part: Modulo an $(n+2a)/(n+2a-4)$ term, this is the expectation of σ^2 from (3.4).

Exercise 3.5. Give a $(1-\alpha)$ HPD region on β based on (3.6).

Exercise 3.6. This regression model can also be used in a predictive sense: For a given $(m, k+1)$ explanatory matrix $\tilde{X}$, the corresponding outcome $\tilde{\mathbf{y}}$ can be inferred through the *predictive distribution* $\pi(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X})$. Show that $\pi(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X})$ is a Gaussian density with mean

$$\begin{aligned}\mathbb{E}[\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X}] &= \mathbb{E}[\mathbb{E}(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}, X, \tilde{X})|\sigma^2, \mathbf{y}, X, \tilde{X}]\\ &= \mathbb{E}[\tilde{X}\beta|\sigma^2, \mathbf{y}, X, \tilde{X}]\\ &= \tilde{X}(M + X^\mathsf{T}X)^{-1}(X^\mathsf{T}X\hat{\beta} + M\tilde{\beta})\end{aligned}$$

and covariance matrix

$$\begin{aligned}\mathbb{V}(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, \tilde{X}) &= \mathbb{E}[\mathbb{V}(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}, X, \tilde{X})|\sigma^2, \mathbf{y}, X, \tilde{X}]\\ &\quad + \mathbb{V}[\mathbb{E}(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}, X, \tilde{X})|\sigma^2, \mathbf{y}, X, \tilde{X}]\\ &= \mathbb{E}[\sigma^2 I_m|\sigma^2, \mathbf{y}, X, \tilde{X}] + \mathbb{V}(\tilde{X}\beta|\sigma^2, \mathbf{y}, X, \tilde{X})\\ &= \sigma^2(I_m + \tilde{X}(M + X^\mathsf{T}X)^{-1}\tilde{X}^\mathsf{T}).\end{aligned}$$

Deduce that

$$\begin{aligned}\tilde{\mathbf{y}}|\mathbf{y}, X, \tilde{X} \sim \mathscr{T}_m\Big(&n+2a, \tilde{X}(M + X^\mathsf{T}X)^{-1}(X^\mathsf{T}X\hat{\beta} + M\tilde{\beta}),\\ &\frac{2b + s^2 + (\tilde{\beta}-\hat{\beta})^\mathsf{T}\left(M^{-1} + (X^\mathsf{T}X)^{-1}\right)^{-1}(\tilde{\beta}-\hat{\beta})}{n+2a}\\ &\times\left\{I_m + \tilde{X}(M + X^\mathsf{T}X)^{-1}\tilde{X}^\mathsf{T}\right\}\Big).\end{aligned}$$

Exercise 3.7. Show that the marginal distribution of $\mathbf{y}$ associated with (3.3) and (3.4) is given by

$$\mathbf{y}|X \sim \mathscr{T}_n\left(2a, X\tilde{\beta}, \frac{b}{a}(I_n + XM^{-1}X^{\mathsf{T}})\right).$$

Exercise 3.8. Given the null hypothesis $H_0 : R\beta = 0$, where R is a (q, p) matrix of rank q, show that the restricted model on $\mathbf{y}$ given X can be represented as

$$\mathbf{y}|\beta_0, \sigma_0^2, X_0 \overset{H_0}{\sim} \mathscr{N}_n\left(X_0\beta_0, \sigma_0^2 I_n\right),$$

where X_0 is an $(n, k-q)$ matrix and β_0 is a $(k-q)$ dimensional vector. (*Hint:* Give the form of X_0 and β_0 in terms of X and β.) Under the hypothesis-specific prior $\beta_0|H_0, \sigma_0^2 \sim \mathscr{N}_{k-q}\left(\tilde{\beta}_0, \sigma^2(M_0)^{-1}\right)$ and $\sigma_0^2|H_0 \sim \mathscr{IG}(a_0, b_0)$, construct the Bayes factor associated with the test of H_0.

Obviously, the choices of the hyperparameters $a, b, \tilde{\beta}$, and M in both (3.3) and (3.4) are important, but they are not always easy in practice. In particular, building prior beliefs on the correlation between the components of β is often difficult. This is one of the reasons why M is frequently chosen as a diagonal matrix or even as a multiple of the identity matrix, $M = I_{k+1}/c$.

⊞ For caterpillar, assume we have no precise prior information about the hyperparameters $\tilde{\beta}$, M, a, and b. For illustrative purposes, we choose $a = 2.1$ and $b = 2$, which correspond to a prior mean and a prior variance of

$$\frac{2}{2.1-1} \approx 1.82 \quad \text{and} \quad \frac{2^2}{(2.1-2)^2(2.1-2)} \approx 33.06$$

on σ^2, respectively. Intuitively, the prior variance is large so that the prior distribution of σ^2 should have a rather weak influence on the outcome. Similarly, we choose $\tilde{\beta} = 0_{k+1}$ and $M = I_{k+1}/c$, that is,

$$\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(0_{k+1}, c\sigma^2 I_{k+1}).$$

As usual, the intuition is that if c is larger, the prior distribution on β should be more diffuse and thus should have less bearing on the outcome. In contradiction with this intuition, Table 3.1 highlights the lasting influence of the factor c on the posterior means of both σ^2 and β_1. The value of c thus has a significant influence on the estimators and even more on the posterior variance, and the Bayes estimates only stabilize for *very large* values of c. In practice, this means that the prior associated with a particular choice of c should not be considered as a weak or pseudo-noninformative prior but, on the opposite, associated with a specific proper prior information. (As a further illustration, Table 3.2 gives the conjugate Bayes estimate of β for $c = 100$.) Although not illustrated here, the dependence on (a, b) is equally strong.

Table 3.1. Dataset caterpillar: Influence of the prior scale c on the Bayes estimates of σ^2 and β_0.

c	$\mathbb{E}(\sigma^2\|\mathbf{y},X)$	$\mathbb{E}(\beta_0\|\mathbf{y},X)$	$\mathbb{V}(\beta_0\|\mathbf{y},X)$
.1	1.0044	0.1251	0.0988
1	0.8541	0.9031	0.7733
10	0.6976	4.7299	3.8991
100	0.5746	9.6626	6.8355
1000	0.5470	10.8476	7.3419

Table 3.2. Dataset caterpillar: Bayes estimates of β for $c = 100$.

β_i	$\mathbb{E}(\beta_i\|\mathbf{y},X)$	$\mathbb{V}(\beta_i\|\mathbf{y},X)$
β_0	9.6626	6.8355
β_1	–0.0041	2×10^{-6}
β_2	–0.0516	0.0004
β_3	0.0418	0.0076
β_4	–1.2633	0.2615
β_5	0.2307	0.0090
β_6	–0.0832	1.9310
β_7	–0.1917	0.8254
β_8	0.1608	0.0046
β_9	–1.2069	0.6127
β_{10}	–0.2567	0.4267

Given that a natural conjugate prior shows its limitations in this setup in the sense that there is some lasting influence of the hyperparameters on at least the posterior variance, a more elaborate noninformative strategy is called for. Section 3.3 concentrates on a noninformative prior analysis. However, we first consider a middle-ground perspective where the prior information is available on β only, adopting the so-called[5] *Zellner's G-prior*, which somehow settles the problem of the choice of M.

3.2.2 Zellner's *G*-Prior

The idea at the core of Zellner's G-prior modeling is to allow the experimenter to introduce information about the location parameter of the regression but to bypass the most difficult aspects of the prior specification, namely the derivation of the prior correlation structure. This structure is fixed in Zellner's proposal since the prior corresponds to

$$\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^{\mathsf{T}}X)^{-1}) \quad \text{and} \quad \pi(\sigma^2|X) \propto \sigma^{-2}\,.$$

Zellner's G-prior thus relies on a (conditional) Gaussian prior for β and an improper (Jeffreys) prior for σ^2. It somehow appears as a data-dependent

[5] Arnold Zellner is a famous Bayesian econometrician from the University of Chicago, who wrote two reference books on Bayesian econometrics (Zellner, 1971, 1984). The "G" in the G-prior comes from the constant G used by Zellner in the prior variance, a constant denoted by c below.

prior through its dependence on X, but this is not really a problem[6] since the *whole* model is conditional on X. The experimenter thus restricts prior determination to the choices of $\tilde{\beta}$ and the constant c. Note that c can be interpreted as a measure of the amount of information available in the prior relative to the sample. For instance, setting $1/c = 0.5$ gives the prior the same weight as 50% of the sample.

↯ Genuine data-dependent priors are not acceptable in a Bayesian analysis because they use the data *twice* and generally miss the basic convergence properties of the Bayes estimators. (See Carlin and Louis, 1996, for a comparative study of the corresponding empirical Bayes estimators.)

With this prior model, the posterior simplifies into

$$\begin{aligned}\pi(\beta,\sigma^2|\mathbf{y},X) &\propto f(\mathbf{y}|\beta,\sigma^2,X)\pi(\beta,\sigma^2|X)\\ &\propto \left(\sigma^2\right)^{-(n/2+1)}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-X\hat{\beta})^{\mathsf{T}}(\mathbf{y}-X\hat{\beta})\right.\\ &\quad\left.-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\beta-\hat{\beta})\right]\left(\sigma^2\right)^{-k/2}\\ &\quad\times\exp\left[-\frac{1}{2c\sigma^2}(\beta-\tilde{\beta})X^{\mathsf{T}}X(\beta-\tilde{\beta})\right],\end{aligned}$$

given that $X^{\mathsf{T}}X$ is used in both the prior and the likelihood. Therefore,

$$\begin{aligned}\beta|\sigma^2,\mathbf{y},X &\sim \mathcal{N}_{k+1}\left(\frac{c}{c+1}(\tilde{\beta}/c+\hat{\beta}),\frac{\sigma^2 c}{c+1}(X^{\mathsf{T}}X)^{-1}\right),\\ \sigma^2|\mathbf{y},X &\sim \mathcal{IG}\left(\frac{n}{2},\frac{s^2}{2}+\frac{1}{2(c+1)}(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\tilde{\beta}-\hat{\beta})\right).\end{aligned}$$

Exercise 3.9. Show that

$$\begin{aligned}\beta|\mathbf{y},X \sim \mathcal{T}_{k+1}\left(n,\frac{c}{c+1}\left(\frac{\tilde{\beta}}{c}+\hat{\beta}\right),\right.\\ \left.\frac{c(s^2+(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)}(X^{\mathsf{T}}X)^{-1}\right).\end{aligned}$$

The Bayes estimates of β and σ^2 are given by

[6] This choice is more problematic when conditioning on X is no longer possible, as for instance when X contains lagged dependent variables (Chapter 7) or endogenous variables.

$$\mathbb{E}[\beta|\mathbf{y}, X] = \frac{1}{c+1}(\tilde{\beta} + c\hat{\beta}) \tag{3.7}$$

and

$$\mathbb{E}[\sigma^2|\mathbf{y}, X] = \frac{s^2 + (\tilde{\beta} - \hat{\beta})^\mathsf{T} X^\mathsf{T} X(\tilde{\beta} - \hat{\beta})/(c+1)}{n-2}\,. \tag{3.8}$$

respectively. Equations (3.7) and (3.8) both highlight the role of c much more clearly than in the case above. When c goes to infinity, the influence of the prior vanishes at speed $1/c$. Moreover,

$$\mathbb{V}[\beta|\mathbf{y}, X] = \frac{c}{c+1}\,\frac{(s^2 + (\tilde{\beta} - \hat{\beta})^\mathsf{T} X^\mathsf{T} X(\tilde{\beta} - \hat{\beta})/(c+1))}{n}(X^\mathsf{T}X)^{-1}\,.$$

The G-priors are also very convenient tools for translating the prior information on β in *virtual sample sizes*. For instance, if we set $c = 1$, this is equivalent to putting the same weight on the prior information and on the sample. In this case,

$$\mathbb{E}(\beta|\mathbf{y}, X) = \frac{\tilde{\beta} + \hat{\beta}}{2}\,,$$

which is simply the average between the prior mean and the maximum likelihood estimator. If, instead, $c = 100$, the prior gets a weight corresponding to 1% of the sample.

⊞ For $\tilde{\beta} = 0_{k+1}$ and $c = 100$, the G-prior estimate of σ^2 for caterpillar is equal to 0.506, while the posterior means and variances of the β_i's are given in Table 3.3. For $\tilde{\beta} = 0_{k+1}$ and $c = 1000$, the G-prior estimate of σ^2 is equal to 0.4899, and the posterior means and variances of the β_i's are given in Table 3.4. Therefore, the G-prior estimates of both σ^2 and β are also sensitive to different (even large) values of c. (Note that while the estimates of β_0 vary less between Table 3.3 and Table 3.4 than between Table 3.1 and Table 3.2, this is due to the fact that the $(0,0)$ term in $(X^\mathsf{T}X)^{-1}$ is approximately 13.) The proper scale of c corresponding to the vanishing influence being data-dependent, it should be addressed via a proper noninformative methodology, as developed in Section 3.3.

As already stressed above (Exercise 3.6), the prediction of $m \geq 1$ future observations from units for which the explanatory variables $\tilde{X}$—but not the outcome variable $\tilde{\mathbf{y}}$—have been observed can be based on the posterior distribution. Obviously, were β and σ^2 exactly known, the m-vector $\tilde{\mathbf{y}}$ would then have a Gaussian distribution with mean $\tilde{X}\beta$ and variance $\sigma^2 I_m$. The *predictive distribution* on $\tilde{\mathbf{y}}$ is defined as the marginal of the joint posterior distribution on $(\tilde{\mathbf{y}}, \beta, \sigma^2)$ and, in the current setting, it can be computed analytically by integrating the conditional posterior predictive distribution $\pi(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X})$ against the posterior $\pi(\sigma^2|\mathbf{y}, X, \tilde{X})$.

Exercise 3.10. Show that $\pi(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X})$ is a Gaussian density.

Table 3.3. Dataset caterpillar: Posterior mean and variance of β for $c = 100$ using Zellner's G-prior.

β_i	$\mathbb{E}(\beta_i\|\mathbf{y}, X)$	$\mathbb{V}(\beta_i\|\mathbf{y}, X)$
β_0	10.8895	6.4094
β_1	–0.0044	2×10^{-6}
β_2	–0.0533	0.0003
β_3	0.0673	0.0068
β_4	–1.2808	0.2175
β_5	0.2293	0.0075
β_6	–0.3532	1.6793
β_7	–0.2351	0.6926
β_8	0.1793	0.0383
β_9	–1.2726	0.5119
β_{10}	–0.4288	0.3696

Table 3.4. Dataset caterpillar: Same legend as Table 3.3 for $c = 1000$.

β_i	$\mathbb{E}(\beta_i\|\mathbf{y}, X)$	$\mathbb{V}(\beta_i\|\mathbf{y}, X)$
β_0	10.9874	6.2604
β_1	–0.0044	2×10^{-6}
β_2	–0.0538	0.0003
β_3	0.0679	0.0066
β_4	–1.2923	0.2125
β_5	0.2314	0.0073
β_6	–0.3564	1.6403
β_7	-0.2372	0.6765
β_8	0.1809	0.0375
β_9	–1.2840	0.5100
β_{10}	–0.4327	0.3670

Conditional on σ^2, the future vector $\tilde{\mathbf{y}}$ of observations has a Gaussian distribution and we can derive its expectation by averaging over β,

$$\begin{aligned}\mathbb{E}[\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X}] &= \mathbb{E}[\mathbb{E}(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}, X, \tilde{X})|\sigma^2, \mathbf{y}, X, \tilde{X}] \\ &= \mathbb{E}[\tilde{X}\beta|\sigma^2, \mathbf{y}, X, \tilde{X}] \\ &= \tilde{X}\frac{\tilde{\beta} + c\hat{\beta}}{c+1}\,,\end{aligned}$$

independently of σ^2. Similarly, we can compute

$$\begin{aligned}\mathbb{V}(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}, X, \tilde{X}) &= \mathbb{E}[\mathbb{V}(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}, X, \tilde{X})|\sigma^2, \mathbf{y}, X, \tilde{X}] \\ &\quad + \mathbb{V}(\mathbb{E}(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}, X, \tilde{X})|\sigma^2, \mathbf{y}, X, \tilde{X}) \\ &= \mathbb{E}[\sigma^2 I_m|\sigma^2, \mathbf{y}, X, \tilde{X}] + \mathbb{V}(\tilde{X}\beta|\sigma^2, \mathbf{y}, X, \tilde{X}) \\ &= \sigma^2\left(I_m + \frac{c}{c+1}\tilde{X}(X^{\mathsf{T}}X)^{-1}\tilde{X}^{\mathsf{T}}\right).\end{aligned}$$

This decomposition of the variance is rather natural: Conditionally on σ^2, the posterior predictive variance has two terms, the first one being $\sigma^2 I_m$, which corresponds to the sampling variation, and the second one being $\sigma^2 \frac{c}{c+1}\tilde{X}(X^{\mathsf{T}}X)^{-1}\tilde{X}^{\mathsf{T}}$, which corresponds to the uncertainty about β.

A standard predictor is the posterior predictive mean, that is,

$$\tilde{X}\frac{\tilde{\beta}+c\hat{\beta}}{c+1}\,,$$

which corresponds to the squared error loss. This representation is once more quite intuitive, being the product of the matrix of explanatory variables $\tilde{X}$ by the Bayes estimate of β.

Exercise 3.11. The true posterior predictive distribution is obtained by integration over the marginal posterior distribution of σ^2. Derive $\pi(\tilde{\mathbf{y}}|\mathbf{y}, X, \tilde{X})$.

Highest posterior density (HPD) regions on subvectors of the parameter β can be derived in a straightforward manner from the marginal posterior distribution of β. For a single parameter, we have for instance

$$\beta_i|\mathbf{y}, X \sim \mathscr{T}_1\left(n, \frac{c}{c+1}\left(\frac{\tilde{\beta}_i}{c}+\hat{\beta}_i\right), \frac{c(s^2+(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)}\omega_{(i,i)}\right)\,,$$

where $\omega_{(i,i)}$ is the (i,i)th element of the matrix $(X^{\mathsf{T}}X)^{-1}$. If we let

$$\tau = \frac{\tilde{\beta}+c\hat{\beta}}{c+1} \quad \text{and} \quad K = \frac{c(s^2+(\tilde{\beta}-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)}(X^{\mathsf{T}}X)^{-1}\,,$$

with components $\kappa_{(i,j)}$, the transform

$$\mathfrak{T}_i = \frac{\beta_i-\tau_i}{\sqrt{\kappa_{(i,i)}}}$$

has a marginal posterior distribution equal to a standard t distribution with n degrees of freedom. A $(1-\alpha)$ HPD interval on β_i is thus given by

$$\left[\tau_i - \sqrt{\kappa_{(i,i)}}F_n^{-1}(1-\alpha/2), \tau_i + \sqrt{\kappa_{(i,i)}}F_n^{-1}(1-\alpha/2)\right]$$

when F_n is the $\mathscr{T}_n$ cdf.

Exercise 3.12. Give a joint $(1-\alpha)$ HPD region on β.

The marginal distribution of $\mathbf{y}$ is once again a multivariate t distribution. In fact, since $\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^{\mathsf{T}}X)^{-1})$, the linear transform of β satisfies

$$X\beta|\sigma^2, X \sim \mathcal{N}_n(X\tilde{\beta}, c\sigma^2 X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}),$$

which implies that

$$\mathbf{y}|\sigma^2, X \sim \mathcal{N}_n(X\tilde{\beta}, \sigma^2(I_n + cX(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}})).$$

Integrating in σ^2 with $\pi(\sigma^2) = 1/\sigma^2$ yields

$$f(\mathbf{y}|X) = (c+1)^{-(k+1)/2}\pi^{-n/2}\Gamma(n/2)\left[\mathbf{y}^{\mathsf{T}}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y} - \frac{1}{c+1}\tilde{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\tilde{\beta}\right]^{-n/2}.$$

Following the general methodology presented in Chapter 2, we can therefore conduct Bayesian testing by constructing closed-form Bayes factors. For instance, if the null hypothesis is $H_0 : R\beta = r$ as in Exercise 3.8, the model under H_0 can be rewritten as

$$\mathbf{y}|\beta^0, \sigma^2, X_0 \overset{H_0}{\sim} \mathcal{N}_n\left(X_0\beta^0, \sigma^2 I_n\right),$$

where β^0 is $(k+1-q)$-dimensional.

Exercise 3.13. Show that the matrix $(I_n + cX(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}})$ has 1 and $c+1$ as eigenvalues. (*Hint:* Show that the eigenvectors associated with $c+1$ are of the form $X\beta$ and that the eigenvectors associated with 1 are those orthogonal to X, i.e. the z's such that $X^{\mathsf{T}}z = 0$.) Deduce that the determinant of the matrix $(I_n + cX(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}})$ is indeed $(c+1)^{(k+1)/2}$.

It is paramount to use the *same* σ^2 in both models. In fact, as discussed first in Section 2.3.3, the ban on improper priors on σ^2 can only be bypassed by imposing a common parameter and hence a common improper prior. Otherwise, the Bayes factor cannot be correctly defined.

Under the prior

$$\beta^0|X_0, \sigma^2 \sim \mathcal{N}_{k+1-q}\left(\tilde{\beta}^0, c_0\sigma^2(X_0^{\mathsf{T}}X_0)^{-1}\right), \tag{3.9}$$

the marginal distribution of $\mathbf{y}$ under H_0 is

$$f(\mathbf{y}|X_0,H_0) = (c+1)^{-(k+1-q)/2}\pi^{-n/2}\Gamma(n/2)\left[\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c_0}{c_0+1}\mathbf{y}^\mathsf{T}X_0(X_0^\mathsf{T}X_0)^{-1}X_0^\mathsf{T}\mathbf{y} - \frac{1}{c_0+1}\tilde{\beta}_0^\mathsf{T}X_0^\mathsf{T}X_0\tilde{\beta}_0\right]^{-n/2},$$

and therefore, the Bayes factor is given in closed form by

$$B_{10}^\pi = \frac{f(\mathbf{y}|X,H_1)}{f(\mathbf{y}|X_0,H_0)} = \frac{(c_0+1)^{(k+1-q)/2}}{(c+1)^{(k+1)/2}} \times \left[\frac{\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c_0}{c_0+1}\mathbf{y}^\mathsf{T}X_0(X_0^\mathsf{T}X_0)^{-1}X_0^\mathsf{T}\mathbf{y} - \frac{1}{c_0+1}\tilde{\beta}_0^\mathsf{T}X_0^\mathsf{T}X_0\tilde{\beta}_0}{\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^\mathsf{T}X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y} - \frac{1}{c+1}\tilde{\beta}^\mathsf{T}X^\mathsf{T}X\tilde{\beta}}\right]^{n/2}.$$

↯ The dependence of the Bayes factor on the pair (c,c_0) cannot be bypassed in the sense that the Bayes factor varies between 0 and ∞ when c_0/c goes from 0 to ∞. This fundamental indeterminacy is related to Section 2.3.3 but can be solved by imposing on (3.9) a consistency constraint: This prior can appear as a projection of the general prior, in which case both $\tilde{\beta}^0$ and $c=c_0$ are uniquely defined.

⊞ For caterpillar, if we want to test $H_0 : \beta_8 = \beta_9 = 0$, using $\tilde{\beta} = 0_{11}$, $\tilde{\beta}^0 = 0_9$, and $c = c_0 = 100$, we obtain $B_{10}^\pi = 0.0165$. Using Jeffreys' scale of evidence, since $\log_{10}(B_{10}^\pi) = -1.78$, the posterior distribution is strongly in favor of H_0.

More generally, using $\tilde{\beta} = 0_{11}$ and $c = 100$, we can produce the following Bayesian regression output, programmed in R, which mimics the standard software regression output: Besides the estimation of the β_i's given by their posterior expectation, we include the Bayes factors B_{10}^i through the log Bayes factors $\log_{10}\left(B_{10}^i\right)$ corresponding to the null hypotheses $H_0 : \beta_i = 0$. (The stars are related to Jeffreys' scale of evidence.)

```
             Estimate  Post. Var.  log10(BF)

(Intercept) 10.8895   6.4094       2.1873 (****)
X1          -0.0044    2e-06       1.1571 (***)
X2          -0.0533   0.0003       0.6667 (**)
X3           0.0673   0.0068      -0.8585
X4          -1.2808   0.2175       0.4726 (*)
X5           0.2293   0.0075       0.3861 (*)
X6          -0.3532   1.6793      -0.9860
X7          -0.2351   0.6926      -0.9848
X8           0.1793   0.0383      -0.8223
X9          -1.2726   0.5119      -0.3461
X10         -0.4288   0.3696      -0.8949

evidence against H0: (****) decisive, (***) strong,
(**) substantial, (*) poor
```

Although these Bayes factors should not be used simultaneously, an informal conclusion is that the most important variables besides the intercept seem to be X_1, X_2, X_4, and X_5.

3.3 Noninformative Prior Analyses

Given the robustness limitations of both solutions developed in Section 3.2 in the case of a complete lack of prior information, we now consider two noninformative solutions that are more elaborate.

3.3.1 Jeffreys' Prior

In this case, Jeffreys' prior is equal to

$$\pi(\beta, \sigma^2|X) \propto \sigma^{-2}\,,$$

that is, a flat prior on $(\beta, \log\sigma^2)$. The corresponding posterior distribution is

$$\begin{aligned}\pi(\beta, \sigma^2|\mathbf{y}, X) &\propto \left(\sigma^{-2}\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-X\hat{\beta})^{\mathsf{T}}(\mathbf{y}-X\hat{\beta})\right.\\ &\qquad \left.-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\beta-\hat{\beta})\right] \times \sigma^{-2}\\ &\propto \left(\sigma^{-2}\right)^{-(k+1)/2} \exp\left[-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\beta-\hat{\beta})\right]\\ &\qquad \times \left(\sigma^{-2}\right)^{-(n-k-1)/2-1} \exp\left[-\frac{1}{2\sigma^2}s^2\right]\,.\end{aligned}$$

We thus deduce the two-stage distribution

$$\begin{aligned}\beta|\sigma^2, \mathbf{y}, X &\sim \mathcal{N}_{k+1}\left(\hat{\beta}, \sigma^2(X^{\mathsf{T}}X)^{-1}\right),\\ \sigma^2|\mathbf{y}, X &\sim \mathcal{IG}((n-k-1)/2, s^2/2).\end{aligned}$$

As in every analysis using an improper prior, one needs to check that the posterior distribution is proper. In this case, $\pi(\beta, \sigma^2|\mathbf{y}, X)$ is proper when $n > k$ and X is of rank $(k+1)$. The former constraint requires that there be at least as many data points as there are parameters in the model, and the latter is obviously necessary for identifiability reasons.

Exercise 3.14. Derive the marginal posterior distribution of β for this model.

Exercise 3.15. Show that the marginal posterior distribution of β_i $(1 \le i \le k)$ is a $\mathcal{T}_1(n-k-1, \hat{\beta}_i, \omega_{(i,i)} s^2/(n-k-1))$ distribution. (*Hint:* Recall that $\omega_{(i,i)} = (X^{\mathsf{T}}X)^{-1}_{(i,i)}$.)

The corresponding estimates of β and σ^2 are given by

$$\mathbb{E}[\beta|\mathbf{y}, X] = \hat{\beta} \quad \text{and} \quad \mathbb{E}[\sigma^2|\mathbf{y}, X] = \frac{s^2}{n-k-3}\,,$$

respectively. Therefore, while the Jeffreys estimate of β is the maximum likelihood estimate, the Jeffreys estimate of σ^2 is larger (and thus more pessimistic) than the maximum likelihood estimate s^2/n and even than the classical estimate $s^2/(n-k-1)$.

The similarity with the frequentist analysis of this model is very strong since the classical $(1-\alpha)$ confidence interval and the Bayesian HPD interval on β_i then coincide, even though they have different interpretations. They are both equal to

$$\left\{\beta_i;\ |\beta_i - \hat{\beta}_i| \le F^{-1}_{n-k}(1-\alpha/2)\sqrt{\omega_{(i,i)} s^2/(n-k-1)}\right\}$$

(see Exercise 3.15).

⊞ For caterpillar, the Bayes estimate of σ^2 is equal to 0.756. Table 3.5 provides the corresponding (marginal)[7] 95% HPD intervals for each component of β. Note that while some of these intervals include the value $\beta_i = 0$, they do not necessarily validate the null hypothesis $H_0 : \beta_i = 0$. This is a major difference from the classical approach, where confidence intervals are dual sets of acceptance regions.

Exercise 3.16. Give the predictive distribution of $\tilde{\mathbf{y}}$, the m dimensional vector corresponding to the (m, k) matrix of explanatory variables $\tilde{X}$.

↯ When using this improper prior distribution, testing point null hypotheses is once more impossible for the reasons advanced in Section 2.3.3.

[7] These intervals should not be interpreted jointly in the sense that the probability that the vector β belongs to the corresponding hypercube $[5, 7, 16.2] \times \cdots \times [-1.7, 0.8]$ is *not* equal to 0.95.

Table 3.5. Dataset caterpillar: 95% HPD intervals for the components of β for $c = 100$.

β_i	HPD Interval
β_0	$[5.7435, 16.2533]$
β_1	$[-0.0071, -0.0018]$
β_2	$[-0.0914, -0.0162]$
β_3	$[-0.1029, 0.2387]$
β_4	$[-2.2618, -0.3255]$
β_5	$[0.0524, 0.4109]$
β_6	$[-3.0466, 2.3330]$
β_7	$[-1.9649, 1.4900]$
β_8	$[-0.2254, 0.5875]$
β_9	$[-2.7704, 0.1997]$
β_{10}	$[-1.6950, 0.8288]$

3.3.2 Zellner's Noninformative *G*-Prior

The warning above is one of the reasons why we now describe an alternative to the Jeffreys prior, based on Zellner's G-priors. The difference from the first-level G-prior construction above is that we now consider c an unknown hyperparameter; that is, we use the same G-prior distribution as in Section 3.2.2 with $\tilde{\beta} = 0_{k+1}$, conditionally on c, and we now introduce a diffuse prior distribution on c,

$$\pi(c) = c^{-1}\mathbb{I}_{\mathbb{N}^*}(c)\,.$$

The corresponding marginal posterior on the parameters of interest is then

$$\begin{aligned}\pi(\beta, \sigma^2|\mathbf{y}, X) &= \int \pi(\beta, \sigma^2|\mathbf{y}, X, c)\pi(c|\mathbf{y}, X)\,\mathrm{d}c \\ &\propto \sum_{c=1}^{\infty} \pi(\beta, \sigma^2|\mathbf{y}, X, c) f(\mathbf{y}|X, c)\pi(c) \\ &\propto \sum_{c=1}^{\infty} \pi(\beta, \sigma^2|\mathbf{y}, X, c) f(\mathbf{y}|X, c)\, c^{-1}\,.\end{aligned}$$

The choice of the integer support is mostly computational, while the Jeffreys-like $1/c$ shape is appropriate for a scale parameter.

The conditional distribution $\pi(\beta, \sigma^2|\mathbf{y}, X, c)$ is given in Section 3.2.2 and

$$f(\mathbf{y}|X, c) \propto (c+1)^{-(k+1)/2}\left[\mathbf{y}^{\mathsf{T}}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}\right]^{-n/2}.$$

While this expression of $\pi(\beta, \sigma^2|\mathbf{y}, X)$ is formally given in a closed form, the infinite summation is a problem in practice. For instance, the Bayes estimates of β and σ^2 are given by

$$\mathbb{E}[\beta|\mathbf{y},X] = \mathbb{E}[\mathbb{E}(\beta|\mathbf{y},X,c)|\mathbf{y},X] = \mathbb{E}[c/(c+1)\hat{\beta})|\mathbf{y},X]$$
$$= \left(\frac{\sum_{c=1}^{\infty} c/(c+1) f(\mathbf{y}|X,c) c^{-1}}{\sum_{c=1}^{\infty} f(\mathbf{y}|X,c) c^{-1}} \right) \hat{\beta} \tag{3.10}$$

and

$$\mathbb{E}[\sigma^2|\mathbf{y},X] = \frac{\sum_{c=1}^{\infty} \dfrac{s^2 + \hat{\beta}^{\mathsf{T}} X^{\mathsf{T}} X \hat{\beta}/(c+1)}{n-2} f(\mathbf{y}|X,c) c^{-1}}{\sum_{c=1}^{\infty} f(\mathbf{y}|X,c) c^{-1}}, \tag{3.11}$$

respectively. Both terms involve infinite summations on c. Note also that the denominator in both cases is the normalizing constant of the posterior. Moreover, the variance of β is obtained as

$$\begin{aligned}
\mathbb{V}[\beta|\mathbf{y},X] &= \mathbb{E}[\mathbb{V}(\beta|\mathbf{y},X,c)|\mathbf{y},X] + \mathbb{V}[\mathbb{E}(\beta|\mathbf{y},X,c)|\mathbf{y},Xx] \\
&= \mathbb{E}\left[c/(n(c+1))(s^2 + \hat{\beta}^{\mathsf{T}}(X^{\mathsf{T}}X)\hat{\beta}/(c+1))(X^{\mathsf{T}}X)^{-1}\right] \\
&\quad + \mathbb{V}[c/(c+1)\hat{\beta}|\mathbf{y},X] \\
&= \left[\frac{\sum_{c=1}^{\infty} f(\mathbf{y}|X,c)/(n(c+1))(s^2 + \hat{\beta}^{\mathsf{T}}(X^{\mathsf{T}}X)\hat{\beta}/(c+1))}{\sum_{c=1}^{\infty} f(\mathbf{y}|X,c) c^{-1}} \right] (X^{\mathsf{T}}X)^{-1} \\
&\quad + \hat{\beta} \left(\frac{\sum_{c=1}^{\infty} (c/(c+1) - \mathbb{E}(c/(c+1)|\mathbf{y},X))^2 f(\mathbf{y}|X,c) c^{-1}}{\sum_{c=1}^{\infty} f(\mathbf{y}|X,c) c^{-1}} \right) \hat{\beta}^{\mathsf{T}},
\end{aligned}$$

a formidable but explicit formula!

Ⓒ Obviously, the infinite summations in (3.10) and (3.11) need to be truncated. In practice, this means that the upper bound on c is increased until the resulting estimates are constant. In our case, we found that truncating at 10^5 was good enough.

⊞ Under this specific prior, the estimate of σ^2 for caterpillar is equal to 0.7732, and Table 3.6 gives the Bayes estimate of β. Note the difference from the maximum likelihood estimate of Figure 3.2, and from the Bayes estimates when c is fixed. In particular, the estimate of σ^2 is now larger, while the estimate of β_0 is closer to 0 than in Table 3.2.

Table 3.6. Dataset caterpillar: Posterior mean and variance of β under the noninformative Zellner's G-prior.

β_i	$\mathbb{E}(\beta_i\|\mathbf{y},X)$	$\mathbb{V}(\beta_i\|\mathbf{y},X)$
β_0	9.2714	9.1164
β_1	–0.0037	2×10^{-6}
β_2	–0.0454	0.0004
β_3	0.0573	0.0086
β_4	–1.0905	0.2901
β_5	0.1953	0.0099
β_6	–0.3008	2.1372
β_7	–0.2002	0.8815
β_8	0.1526	0.0490
β_9	–1.0835	0.6643
β_{10}	–0.3651	0.4716

Exercise 3.17. When using the prior distribution $\pi(c) = 1/c^2$, compare the results with those of Table 3.6.

An important point with this approach is that the marginal distribution of the dataset is available in closed form (modulo the infinite summation), namely,

$$f(\mathbf{y}|X) \propto \sum_{i=1}^{\infty} c^{-1}(c+1)^{-(k+1)/2}\left[\mathbf{y}^{\mathsf{T}}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}\right]^{-n/2}.$$

It is therefore possible to produce a Bayes regression output, just as in the first-level prior case. This is one additional reason why this noninformative prior is introduced. For instance, if the null hypothesis is $H_0 : R\beta = r$ as in Exercise 3.8, the model under H_0 can be rewritten as

$$\mathbf{y}|\beta^0,\sigma^2,X_0 \overset{H_0}{\sim} \mathcal{N}_n\left(X_0\beta^0,\sigma^2 I_n\right)$$

where β^0 is $(k+1-q)$ dimensional. Under the prior

$$\beta^0|X_0,\sigma^2,c \sim \mathcal{N}_{k+1-q}\left(0_{k+1-q},c\sigma^2(X_0^{\mathsf{T}}X_0)^{-1}\right) \tag{3.12}$$

and $\pi(c) = 1/c$, the marginal distribution of $\mathbf{y}$ under H_0 is

$$f(\mathbf{y}|X_0,H_0) \propto \sum_{c=1}^{\infty}(c+1)^{-(k+1-q)/2}\left[\mathbf{y}^{\mathsf{T}}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^{\mathsf{T}}X_0(X_0^{\mathsf{T}}X_0)^{-1}X_0^{\mathsf{T}}\mathbf{y}\right]^{-n/2},$$

and therefore, the Bayes factor given by

$$B_{10}^{\pi} = \frac{f(\mathbf{y}|X)}{f(\mathbf{y}|X_0,H_0)}$$

can be computed.

⊞ If we test $H_0 : \beta_8 = \beta_9 = 0$ on caterpillar, we obtain $B^{\pi}_{10} = 0.1628$ (which is ten times larger than in the first-level case). Using Jeffreys' scale of evidence, $\log_{10}(B^{\pi}_{10}) = -0.7884$ still clearly favors H_0.

Once more, an R-based Bayes regression output can be produced, as follows:

```
             Estimate  Post. Var.  log10(BF)

(Intercept)  9.2714   9.1164       1.4205 (***)
X1          -0.0037    2e-06       0.8502 (**)
X2          -0.0454   0.0004       0.5664 (**)
X3           0.0573   0.0086      -0.3609
X4          -1.0905   0.2901       0.4520 (*)
X5           0.1953   0.0099       0.4007 (*)
X6          -0.3008   2.1372      -0.4412
X7          -0.2002   0.8815      -0.4404
X8           0.1526   0.0490      -0.3383
X9          -1.0835   0.6643      -0.0424
X10         -0.3651   0.4716      -0.3838

evidence against H0: (****) decisive, (***) strong,
(**) substantial, (*) poor
```

In that case, the conclusion is the same as in the informative case.

Exercise 3.18. Show that both series (3.10) and (3.11) converge.

Exercise 3.19. Give the predictive distribution of $\tilde{\mathbf{y}}$, the m-dimensional vector corresponding to the (m, k) matrix of explanatory variables $\tilde{X}$.

3.4 Markov Chain Monte Carlo Methods

Given the complexity of most models encountered in Bayesian modeling, standard simulation methods are not a sufficiently versatile solution. We now present the rudiments of a technique that emerged in the late 1980s as the core of Bayesian computing and that has since then revolutionized the field.

This technique is based on *Markov chains*, but we will not make many incursions into the theory of Markov chains (see Meyn and Tweedie, 1993), focusing rather on the practical implementation of these algorithms and trusting that the underlying theory is sound enough to validate them (Robert and Casella, 2004). At this point, it is sufficient to recall that a Markov chain $(\mathbf{x}_t)_{t\in\mathbb{N}}$ is a sequence of dependent random vectors whose dependence on the past values $\mathbf{x}_0, \ldots, \mathbf{x}_{t-1}$ stops at the value immediately before, $\mathbf{x}_{t-1}$, and that

is entirely defined by its *kernel*—that is, the conditional distribution of $\mathbf{x}_t$ given $\mathbf{x}_{t-1}$.

The central idea behind these new methods, called *Markov chain Monte Carlo* (MCMC) algorithms, is that, to simulate from a distribution π (for instance, the posterior distribution), it is actually sufficient to produce a Markov chain $(\mathbf{x}_t)_{t\in\mathbb{N}}$ whose *stationary distribution* is π: If $\mathbf{x}_t$ is marginally distributed according to π, then $\mathbf{x}_{t+1}$ is also marginally distributed according to π. If an algorithm that generates such a chain can be constructed, the ergodic theorem guarantees that, in almost all settings, the average

$$\frac{1}{T}\sum_{t=1}^{T} g(\mathbf{x}_t)$$

converges to $\mathbb{E}_\pi[g(\mathbf{x})]$, no matter what the starting value.[8]

More informally, this property means that, for large enough t, $\mathbf{x}_t$ is approximately distributed from π and can thus be used like the output from a more standard simulation algorithm (even though one must take care of the correlations between the $\mathbf{x}_t$'s created by the Markovian structure). For integral approximation purposes, the difference from regular Monte Carlo approximations is that the variance structure of the estimator is more complex because of the Markovian dependence. (These methods being central to the cases studied from this stage onward, we hope that the reader will become sufficiently familiar with them before the end of the book!) In this chapter, we detail a particular type of MCMC algorithm, the Gibbs sampler, that is currently sufficient for our needs. (The next chapter will introduce a more universal type of algorithm.)

3.4.1 Conditionals

A first remark that motivates the use of the Gibbs sampler[9] is that, within structures such as

$$\pi(x_1) = \int \pi_1(x_1|x_2)\tilde{\pi}(x_2)\,\mathrm{d}x_2\,, \tag{3.13}$$

to simulate from the joint distribution

$$\pi_1(x_1|x_2)\tilde{\pi}(x_2) \tag{3.14}$$

automatically produces (marginal) simulation from $\pi(x_1)$. Therefore, in settings where (3.13) holds, it is not necessary to simulate from (3.13) when one can jointly simulate (x_1, x_2) from (3.14).

[8] In probabilistic terms, if the Markov chains produced by these algorithms are irreducible, then these chains are both positive recurrent with stationary distribution π and ergodic, that is, asymptotically independent of the starting value $\mathbf{x}_0$.

[9] In the literature, both the denominations Gibbs *sampler* and Gibbs *sampling* can be found. In this book, we will use Gibbs sampling for the simulation technique and Gibbs sampler for the simulation algorithm.

Example 3.1. Consider $(x_1, x_2) \in \mathbb{N} \times [0,1]$ distributed from the density

$$\pi(x_1, x_2) \propto \binom{n}{x_1} x_2^{x_1+\alpha-1}(1-x_2)^{n-x_1+\beta-1}\,.$$

This is a joint distribution where

$$x_1|x_2 \sim \mathscr{B}(n, x_2) \quad \text{and} \quad x_2|\alpha, \beta \sim \mathscr{B}e(\alpha, \beta)\,.$$

Therefore, although

$$\pi(x_1) = \binom{n}{x_1} \frac{B(\alpha + x_1, \beta + n - x_1)}{B(\alpha, \beta)}$$

is available in closed form as the *beta-binomial distribution*, it is unnecessary to work with this marginal when one can simulate an iid sample $(x_1^{(i)}, x_2^{(i)})$ $(i = 1, \ldots, N)$ as

$$x_2^{(i)} \sim \mathscr{B}e(\alpha, \beta) \text{ and } x_1^{(i)} \sim \mathscr{B}(n, x_2^{(i)})\,.$$

Integrals such as $\mathbb{E}_\pi[x_1/(x_1+1)]$ can then be approximated by

$$\frac{1}{N}\sum_{i=1}^{N} \frac{x_1^{(i)}}{x_1^{(i)}+1}\,,$$

using a regular Monte Carlo approach. ◀

Unfortunately, even when one works with a representation such as (3.13) that is naturally associated with the original model, it is often the case that the mixing density $\tilde{\pi}(x_2)$ itself is neither available in closed form nor amenable to simulation, but rather only the joint density $\pi_1(x_1|x_2)\tilde{\pi}(x_2)$ is manageable. However, both *conditional posterior distributions*,

$$\pi_1(x_1|x_2) \text{ and } \pi_2(x_2|x_1)\,,$$

can often be simulated, and the following method takes full advantage of this feature.

3.4.2 Two-Stage Gibbs Sampler

The availability of both conditionals of (3.13) in terms of simulation can be exploited to build a transition kernel and a corresponding Markov chain, somewhat analogous to the derivation of the maximum of a multivariate function via an iterative device that successively maximizes the function in each of its arguments until a fixed point is reached.

The corresponding Markov kernel is built by simulating successively from each conditional distribution, with the conditioning variable being updated

on the run. It is called the *two-stage Gibbs sampler* or sometimes the *data augmentation* algorithm, although both terms are rather misleading.[10]

ALGORITHM 3.1. TWO-STAGE GIBBS SAMPLER
Initialization: Start with an arbitrary value $x_2^{(0)}$.
Iteration t: Given $x_2^{(t-1)}$, generate
1. $x_1^{(t)}$ according to $\pi_1(x_1|x_2^{(t-1)})$,
2. $x_2^{(t)}$ according to $\pi_2(x_2|x_1^{(t)})$.

Note that, in the second step of the algorithm, $x_2^{(t)}$ is generated conditional on $x_1 = x_1^{(t)}$, not $x_1^{(t-1)}$. The validation of this algorithm is that, for both generations, π is a stationary distribution. Therefore, the limiting distribution of the chain $(x_1^{(t)}, x_2^{(t)})_t$ is π if the chain is *irreducible*; that is, if it can reach any region in the support of π in a finite number of steps. (Note that there is a difference between *stationary* distribution and *limiting* distribution only when the chain is not ergodic, as shown in Exercise 3.20.)

Exercise 3.20. If (x_1, x_2) is distributed from the uniform distribution on

$$\{(x_1, x_2);\ (x_1 - 1)^2 + (x_2 - 1)^2 \le 1\} \cup \{(x_1, x_2);\ (x_1 + 1)^2 + (x_2 + 1)^2 \le 1\},$$

show that the Gibbs sampler does not produce an irreducible chain. For this distribution, find an alternative Gibbs sampler that works. (*Hint*: Consider a rotation of the coordinate axes.)

Exercise 3.21. If a joint density $g(y_1, y_2)$ corresponds to the conditional distributions $g_1(y_1|y_2)$ and $g_2(y_2|y_1)$, show that it is given by

$$g(y_1, y_2) = \frac{g_2(y_2|y_1)}{\int g_2(v|y_1)/g_1(y_1|v)\, \mathrm{d}v}.$$

© The practice of Gibbs sampling involves solving two levels of difficulties: The first level lies in deriving an efficient decomposition of the joint distribution into simulable conditionals and the second one in deciding when to stop the algorithm. Evaluating the efficiency of the decomposition includes assessing the ease of simulation from both conditionals and the level of correlation between the $\mathbf{x}^{(t)}$'s, as well as the *mixing* behavior of the chain, that is, its ability to explore the support of π sufficiently fast. While deciding whether or

[10] Gibbs sampling got its name from *Gibbs fields*, used in image analysis, when Geman and Geman (1984) proposed an early version of this algorithm, while data augmentation refers to Tanner's (1996) special use of this algorithm in missing-data settings, as seen in Chapter 6.

not a given conditional can be simulated is easy enough, it is not always possible to find a manageable conditional, and necessary alternatives such as the *Metropolis–Hastings algorithm* will be described in the following chapters.

Choosing a stopping rule also relates to the mixing performances of the algorithm, as well as to its ability to approximate posterior expectations under π. Many indicators have been proposed in the literature (see Robert and Casella, 2004, Chapter 12) to signify achievement (or lack) of convergence, although none of these is foolproof. In the easiest cases, the lack of convergence is blatant and can be seen on the raw plot of the sequence of $\mathbf{x}^{(t)}$'s, while, in other cases, the Gibbs sampler explores very satisfactorily one mode of the posterior distribution but fails altogether to visit the *other* modes of the posterior: We will encounter such cases in Chapter 6 with mixtures of distributions. Throughout this chapter and the following ones, we will see ways of implementing these recommendations in practice.

Example 3.2. Consider the posterior distribution derived in Example 2.1, for $n = 2$ observations,

$$\pi(\mu|\mathcal{D}) \propto \frac{e^{-\mu^2/20}}{(1+(x_1-\mu)^2)(1+(x_2-\mu)^2)}\,.$$

Even though this is a univariate distribution, it can still be processed by a Gibbs sampler through a data augmentation step. In fact, since

$$\frac{1}{1+(x_i-\mu)^2} = \int_0^\infty e^{-\omega_i[1+(x_i-\mu)^2]}\,\mathrm{d}\omega_i\,,$$

we can define $\boldsymbol{\omega} = (\omega_1, \omega_2)$ and envision $\pi(\mu|\mathcal{D})$ as the marginal distribution of

$$\pi(\mu, \boldsymbol{\omega}|\mathcal{D}) \propto e^{-\mu^2/20} \times \prod_{i=1}^{2} e^{-\omega_i[1+(x_i-\mu)^2]}\,.$$

For this multivariate distribution, a corresponding Gibbs sampler is associated with the following two steps:

1. Generate $\mu^{(t)} \sim \pi(\mu|\boldsymbol{\omega}^{(t-1)}, \mathcal{D})$.
2. Generate $\boldsymbol{\omega}^{(t)} \sim \pi(\boldsymbol{\omega}|\mu^{(t)}, \mathcal{D})$.

The second step is straightforward: The ω_i's are conditionally independent and distributed as $\mathcal{E}xp(1+(x_i-\mu^{(t)})^2)$. The first step is also well-defined since $\pi(\mu|\boldsymbol{\omega}, \mathcal{D})$ is a normal distribution with mean $\sum_i \omega_i x_i/(\sum_i \omega_i + 1/20)$ and variance $1/(2\sum_i \omega_i + 1/10)$. The corresponding R program then simplifies into two single lines

```
> mu = rnorm(1,sum(x*omega)/sum(omega+.05),
+ sqrt(1/(.05+2*sum(omega))))
> omega = rexp(2,1+(x-mu)^2)
```

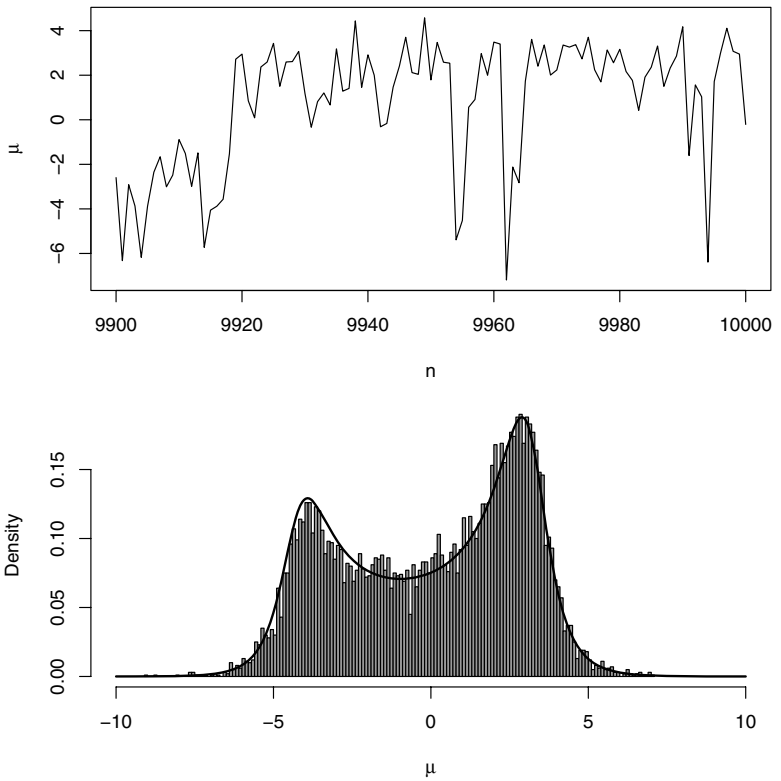


Fig. 3.3. (top) Last 100 iterations of the chain $(\mu^{(t)})$; (bottom) histogram of the chain $(\mu^{(t)})$ and comparison with the target density for $10,000$ iterations.

and the output of the simulation is represented in Figure 3.3, with a very satisfying fit between the histogram of the simulated values and the target. A detailed zoom on the last 100 iterations shows how the chain $(\mu^{(t)})$ moves in the real space, alternatively visiting each mode of the target. ◀

When running a Gibbs sampler, the number of iterations should never be fixed in advance: It is usually impossible to predict the performances of a given sampler before producing a corresponding chain. Deciding on the length of an MCMC run is therefore a sequential process where output behaviors are examined after pilot runs and new simulations (or new samplers) are chosen on the basis of these pilot runs.

Exercise 3.22. Check that the starting value of μ in the setting of Example 3.2 has no influence on the output of the Gibbs sampler above after $N = 1000$ iterations.

3.4.3 The General Gibbs Sampler

For a joint distribution $\pi(x_1, \ldots, x_p)$ with full conditionals $\pi_1, \ldots, \pi_p$ where π_j is the distribution of x_j conditional on $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$, the Gibbs sampler simulates successively from all conditionals, modifying one component of $\mathbf{x}$ at a time. The corresponding algorithmic representation is given in Algorithm 3.2.

ALGORITHM 3.2. GIBBS SAMPLER

Initialization: Start with an arbitrary value $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_p^{(0)})$.

Iteration t: Given $(x_1^{(t-1)}, \ldots, x_p^{(t-1)})$, generate

1. $x_1^{(t)}$ according to $\pi_1(x_1 | x_2^{(t-1)}, \ldots, x_p^{(t-1)})$,
2. $x_2^{(t)}$ according to $\pi_2(x_2 | x_1^{(t)}, x_3^{(t-1)}, \ldots, x_p^{(t-1)})$,

$\vdots$

p. $x_p^{(t)}$ according to $\pi_p(x_p | x_1^{(t)}, \ldots, x_{p-1}^{(t)})$.

Quite logically, the validation of this generalization of Algorithm 3.1 is the same: For each of the p steps of the tth iteration, the joint distribution $\pi(\mathbf{x})$ is stationary. Under the same restriction on the irreducibility of the chain, it also converges to π for every possible starting value. Note that the order of simulation of the components of $\mathbf{x}$ can be modified at each iteration, either deterministically or randomly, without putting the validity of the algorithm in jeopardy.

The two-stage Gibbs sampler naturally appears as a special case of Algorithm 3.2 for $p = 2$. It is, however, endowed with higher theoretical properties, as detailed in Robert and Casella (2004, Chapter 9). In particular, subchains $(x_1^{(t)})_{t \in \mathbb{N}}$ and $(x_2^{(t)})_{t \in \mathbb{N}}$ are both also Markov chains, which is not the case in general for the subchains $(x_i^{(t)})_{t \in \mathbb{N}}$ generated by Algorithm 3.2.

To conclude this section, let us stress that the impact of MCMC on Bayesian statistics has been considerable. Since the 1990s, which saw the emergence of MCMC methods in the statistical community, the occurrence of Bayesian methods in applied statistics has greatly increased, and the frontier between Bayesian and "classical" statistics is now so fuzzy that in some fields, it has completely disappeared. From a Bayesian point of view, the access to

far more advanced computational means has induced a radical modification of the way people work with models and prior assumptions. In particular, it has opened the way to process much more complex structures, such as graphical models and latent variable models (see Chapter 6). It has also freed inference by opening for good the possibility of Bayesian model choice (see, e.g., Robert, 2001, Chapter 7). This expansion is much more visible among academics than among applied statisticians, though, given that the use of the MCMC technology requires some "hard" thinking to process every new problem. The availability of specific software such as BUGS,[11] developed by the Medical Research Council (MRC) Unit in Cambridge (England), has nonetheless given access to MCMC techniques to a wider community, starting with the medical field. New modules in R and other languages are also helping to bridge the gap.

3.5 Variable Selection

3.5.1 Decisional Setting

In an ideal world, when building a regression model, we should include all relevant pieces of information, which in the regression context means including all predictor variables that might possibly help in explaining $\mathbf{y}$. However, there are obvious drawbacks to the advice of increasing the number of explanatory variables. For one thing, in noninformative settings, this eventually clashes with the constraint $k < n$. For another, using a huge number of explanatory variables leaves little information available to obtain precise estimators. In other words, when we increase the explanatory scope of the regression model, we do not necessarily increase its explanatory power because it gets harder and harder to estimate the coefficients.[12] It is thus important to be able to decide which variables within a large pool of potential explanatory variables should be kept in a model that balances good explanatory power with good estimation performances.

This is truly a *decision* problem in that all potential models have to be considered in parallel against a criterion that ranks them. This variable-selection problem can be formalized as follows. We consider a dependent random variable y and a set of k potential explanatory variables. We assume that every subset of q explanatory variables can constitute a proper set of explanatory

[11] BUGS stands for Bayesian inference using Gibbs sampling. There now exists an interface for R users.

[12] This phenomenon is related to the *principle of parsimony*, also called *Occam's razor*, which states that, among two models with similar explanatory powers, the simplest one should always be preferred. It is also related to the *learning curve effect* found in information theory and neural networks, where the performances of a model increase on the learning dataset but decrease on a testing dataset as its complexity increases.

variables for the regression of y. In order to include the case $q = 0$, we assume that the intercept (that is, the constant variable) is included in every model. There are thus 2^k models in competition. Note that the variable-selection problem can alternatively be seen as a two-stage estimation setting where first we estimate the indicator of the model and second we estimate the corresponding parameters.

Each of the 2^k models $\mathfrak{M}_\gamma$ is associated with a binary indicator vector $\gamma \in \Gamma = \{0,1\}^k$, where $\gamma_i = 1$ means that the variable x_i is included in the model $\mathfrak{M}_\gamma$. This notation is quite handy since γ=1010010 clearly indicates which explanatory variables are in and which are not. We also use the notation

$$q_\gamma = \mathbf{1}_n^\mathsf{T}\gamma\,, \quad t_1(\gamma)\,, \quad \text{and} \quad t_0(\gamma)$$

for the number of variables included in the model $\mathfrak{M}_\gamma$, the indices of the variables included in the model $\mathfrak{M}_\gamma$, and the indices of the variables not included in the model $\mathfrak{M}_\gamma$, respectively. Similarly, for $\beta \in \mathbb{R}^{k+1}$ and X, we define β_γ as the subvector

$$\beta_\gamma = \left(\beta_0, (\beta_i)_{i\in t_1(\gamma)}\right)$$

and X_γ as the submatrix of X, where only the column $\mathbf{1}_n$ and the columns in $t_1(\gamma)$ have been left. The model $\mathfrak{M}_\gamma$ is thus defined as

$$\mathbf{y}|\gamma, \beta_\gamma, \sigma^2, X \sim \mathcal{N}_n\left(X_\gamma\beta_\gamma, \sigma^2 I_n\right)\,,$$

where $\beta_\gamma \in \mathbb{R}^{q_\gamma+1}$ and $\sigma^2 \in \mathbb{R}_+^*$ are the unknown parameters.

Once more, and somehow in contradiction to our basic belief that different models should enjoy completely different parameters, we are compelled to use σ^2 as the variance term common to *all models*. Although this is more of a mathematical trick than a true modeling reason, the independence of σ^2 and γ allows the simultaneous use of Bayes factors and an improper prior.

3.5.2 First-Level *G*-Prior Distribution

Because so many models are in competition and thus present in the global model all at once, we cannot expect a practitioner to specify a prior on every $\mathfrak{M}_\gamma$ in a completely subjective and autonomous manner. We thus now proceed to derive *all* priors from a single global prior associated with the so-called *full model* that corresponds to γ=1$\cdots$1. The argument goes as follows:

(i). For the full model, we use Zellner's G-prior as defined in Section 3.2.2,

$$\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^\mathsf{T}X)^{-1}) \quad \text{and} \quad \sigma^2 \sim \pi(\sigma^2|X) = \sigma^{-2}\,.$$

(ii). For each model $\mathfrak{M}_\gamma$, the prior distribution of β_γ conditional on σ^2 is fixed as

$$\beta_\gamma|\gamma,\sigma^2 \sim \mathcal{N}_{q_\gamma+1}\left(\tilde{\beta}_\gamma, c\sigma^2\left(X_\gamma^{\mathsf{T}}X_\gamma\right)^{-1}\right),$$

where $\tilde{\beta}_\gamma = \left(X_\gamma^T X_\gamma\right)^{-1} X_\gamma^T \tilde{\beta}$ and we use the same prior on σ^2. Thus, the joint prior for model $\mathfrak{M}_\gamma$ is the improper prior

$$\pi(\beta_\gamma,\sigma^2|\gamma) \propto \left(\sigma^2\right)^{-(q_\gamma+1)/2-1} \exp\left[-\frac{1}{2(c\sigma^2)}\left(\beta_\gamma-\tilde{\beta}_\gamma\right)^{\mathsf{T}} \left(X_\gamma^{\mathsf{T}}X_\gamma\right)\left(\beta_\gamma-\tilde{\beta}_\gamma\right)\right].$$

↯ This distribution is conditional on γ; in particular, this implies that, while the variance term σ^2 is common to all models, its distribution conditional on γ varies with γ.

Although there are many possible ways of defining the prior on the model index[13] γ, we opt for the uniform prior $\pi(\gamma|X) = 2^{-k}$. The posterior distribution of γ (that is, the distribution of γ given $(\mathbf{y}, X)$) is central to the variable-selection methodology since it is proportional to the marginal density of $\mathbf{y}$ in $\mathfrak{M}_\gamma$. In addition, for prediction purposes, the prediction distribution can be obtained by averaging over all models, the weights being the model probabilities (this is called *model averaging*, as detailed in Section 6.7.3).

The posterior distribution of γ is

$$\begin{aligned}\pi(\gamma|\mathbf{y},X) &\propto f(\mathbf{y}|\gamma,X)\pi(\gamma|X) \propto f(\mathbf{y}|\gamma,X)\\ &= \int\left(\int f(\mathbf{y}|\gamma,\beta,\sigma^2,X)\pi(\beta|\gamma,\sigma^2,X)\,\mathrm{d}\beta\right)\pi(\sigma^2|X)\,\mathrm{d}\sigma^2\,.\end{aligned}$$

Since

$$\begin{aligned}f(\mathbf{y}|\gamma,\sigma^2,X) &= \int f(\mathbf{y}|\gamma,\beta,\sigma^2)\pi(\beta|\gamma,\sigma^2)\,\mathrm{d}\beta\\ &= (c+1)^{-(q_\gamma+1)/2}(2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^{\mathsf{T}}\mathbf{y}\right.\\ &\quad\left.+\frac{1}{2\sigma^2(c+1)}\left\{c\mathbf{y}^{\mathsf{T}}X_\gamma\left(X_\gamma^{\mathsf{T}}X_\gamma\right)^{-1}X_\gamma^{\mathsf{T}}\mathbf{y}-\tilde{\beta}_\gamma^{\mathsf{T}}X_\gamma^{\mathsf{T}}X_\gamma\tilde{\beta}_\gamma\right\}\right),\end{aligned}$$

this posterior density satisfies

$$\begin{aligned}\pi(\gamma|\mathbf{y},X) \propto (c+1)^{-(q_\gamma+1)/2}&\left[\mathbf{y}^{\mathsf{T}}\mathbf{y}-\frac{c}{c+1}\mathbf{y}^{\mathsf{T}}X_\gamma\left(X_\gamma^{\mathsf{T}}X_\gamma\right)^{-1}X_\gamma^{\mathsf{T}}\mathbf{y}\right.\\ &\left.-\frac{1}{c+1}\tilde{\beta}_\gamma^{\mathsf{T}}X_\gamma^{\mathsf{T}}X_\gamma\tilde{\beta}_\gamma\right]^{-n/2}.\end{aligned}$$

[13] For instance, one could use instead a uniform prior on q_γ or a more parsimonious prior such as $\pi(\gamma|X) = 1/q_\gamma$.

Given that, under our prior assumptions, the prior probabilities of all models are identical, the marginal posterior $\pi(\gamma|\mathbf{y}, X)$ can also be used to compute Bayes factors in the sense that $\pi(\gamma|\mathbf{y}, X)$ is proportional to the marginal density $f(\mathbf{y}|X, \gamma)$.

⊞ For caterpillar, we set $\tilde{\beta} = 0_{11}$ and $c = 100$. The models corresponding to the top 20 posterior probabilities are provided in Table 3.7. In a basic $0-1$ decision setup, we would choose the model $\mathfrak{M}_\gamma$ with the highest posterior probability—that is, the model such that $t_1(\gamma) = (1, 2, 4, 5)$—which corresponds to the variables

- altitude,
- slope,
- height of the tree sampled in the center of the area, and
- diameter of the tree sampled in the center of the area.

The selected model corresponds to the five variables identified in the R output at the end of Section 3.2.2, but, interestingly, while the following models all contain the intercept by construction, the explanatory variables x_1 and x_2, which have the most stars in this R analysis, are not (but almost) always included, and x_9, which was excluded in the marginal significance analysis, appears in many of the selected models (and even single-handedly in one case). Note also that (for what it is worth!) a frequentist analysis based on the BIC information criterion leads to the same model.

3.5.3 Noninformative Prior Distribution

Once again, we stress that conventional improper priors do get in the way when comparing models (since this is a special case of a point null test where $H_0^i : \beta_i = 0$ for $i \in t_0(\gamma)$). This problem can somehow be sidestepped by using a *common* improper prior on nuisance parameters that are used over all models. (As pointed out earlier, this goes against our paradigm, but this trick is the most acceptable in the battery of solutions that have been proposed to overcome this difficulty, see Robert, 2001, Chapter 5.) This is the reason why we propose to use again the hierarchical version of the G-prior, with $\tilde{\beta} = 0_{k+1}$ and the hierarchical diffuse prior distribution on c, $\pi(c) \propto c^{-1}\mathbb{I}_{\mathbb{N}^*}(c)$.

⊞ The 20 most likely models and their posterior probabilities are given in Table 3.8. Note that we chose the same model as in the first-level case but that the posterior probabilities of the most likely models are much lower than in Table 3.7. Therefore, the top model is not so strongly supported as in the informative case. Another remark of interest is that, under this noninformative prior, models with a larger number of variables are favored. In fact, the ranking of models is very similar to Table 3.7, except for the exclusion of models with a small number of covariates: $t_1(\gamma) = (1, 9)$, $t_1(\gamma) = (1, 6, 9)$ and $t_1(\gamma) = (9)$ are not preserved, while $t_1(\gamma) = (1, 4, 5)$, and $t_1(\gamma) = (1, 2, 9)$ get a lower ranking.

Table 3.7. Dataset caterpillar: Most likely models ordered by decreasing posterior probabilities under Zellner's G-prior ($c = 100$).

$t_1(\gamma)$	$\pi(\gamma|\mathbf{y}, X)$
0, 1, 2, 4, 5	0.2316
0, 1, 2, 4, 5, 9	0.0374
0, 1, 9	0.0344
0, 1, 2, 4, 5, 10	0.0328
0, 1, 4, 5	0.0306
0, 1, 2, 9	0.0250
0, 1, 2, 4, 5, 7	0.0241
0, 1, 2, 4, 5,8	0.0238
0, 1, 2, 4, 5, 6	0.0237
0,1, 2, 3, 4, 5	0.0232
0, 1, 6, 9	0.0146
0, 1, 2, 3, 9	0.0145
0, 9	0.0143
0, 1, 2, 6, 9	0.0135
0, 1, 4, 5, 9	0.0128
0, 1, 3, 9	0.0117
0, 1, 2, 8	0.0115
0, 1, 8	0.0095
0, 1, 2, 3, 4, 5, 9	0.0090
0, 1, 2, 4, 5, 6, 9	0.0090

Exercise 3.23. Show that

$$\pi(\gamma|\mathbf{y}, X) \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q_\gamma+1)/2} \left[\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{c}{c+1}\mathbf{y}^\mathsf{T} X_\gamma \left(X_\gamma^\mathsf{T} X_\gamma\right)^{-1} X_\gamma^\mathsf{T}\mathbf{y}\right]^{-n/2} \tag{3.15}$$

and that the series converges. If $\pi(c) \propto c^{-\alpha}$, find which values of α lead to a proper posterior.

3.5.4 A Stochastic Search for the Most Likely Model

When the number k of variables gets large, it is impossible to compute the posterior probabilities for the whole series of 2^k models. We then need a tailored algorithm that samples from $\pi(\gamma|\mathbf{y}, X)$ and thus selects the most likely models, without computing first all the values of $\pi(\gamma|\mathbf{y}, X)$. This can be done rather naturally by Gibbs sampling, given the availability of the full conditional posterior probabilities of the γ_i's.

Table 3.8. Dataset caterpillar: 20 most likely models under Zellner's noninformative G-prior.

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y}, X)$
0, 1, 2, 4, 5	0.0929
0, 1, 2, 4, 5, 9	0.0325
0, 1, 2, 4, 5, 10	0.0295
0, 1, 2, 4, 5, 7	0.0231
0, 1, 2, 4, 5, 8	0.0228
0, 1, 2, 4, 5, 6	0.0228
0, 1, 2, 3, 4, 5	0.0224
0, 1, 2, 3, 4, 5, 9	0.0167
0, 1, 2, 4, 5, 6, 9	0.0167
0, 1, 2, 4, 5, 8, 9	0.0137
0, 1, 4, 5	0.0110
0, 1, 2, 4, 5, 9, 10	0.0100
0, 1, 2, 3, 9	0.0097
0, 1, 2, 9	0.0093
0, 1, 2, 4, 5, 7, 9	0.0092
0, 1, 2, 6, 9	0.0092
0, 1, 4, 5, 9	0.0087
0, 1, 2, 3, 4, 5, 10	0.0079
0, 1, 2, 4, 5, 8, 10	0.0079
0, 1, 2, 4, 5, 7, 10	0.0079

Indeed, if γ_{-i} $(1 \le i \le k)$ is the vector $(\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_k)$, the full conditional distribution $\pi(\gamma_i|\mathbf{y}, \gamma_{-i}, X)$ of γ_i is proportional to $\pi(\gamma|\mathbf{y}, X)$ and can be evaluated in both $\gamma_i = 0$ and $\gamma_i = 1$ since these are the only possible values of γ_i.

Algorithm 3.3. Gibbs Sampler for Variable Selection

Initialization: Draw γ^0 from the uniform distribution on Γ.

Iteration t: Given $(\gamma_1^{(t-1)}, \ldots, \gamma_k^{(t-1)})$, generate

1. $\gamma_1^{(t)}$ according to $\pi(\gamma_1|\mathbf{y}, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)}, X)$,
2. $\gamma_2^{(t)}$ according to $\pi(\gamma_2|\mathbf{y}, \gamma_1^{(t)}, \gamma_3^{(t-1)}, \ldots, \gamma_k^{(t-1)}, X)$,

$\vdots$

p. $\gamma_k^{(t)}$ according to $\pi(\gamma_k|\mathbf{y}, \gamma_1^{(t)}, \ldots, \gamma_{k-1}^{(t)}, X)$.

After a large number of iterations (that is, when the sampler is supposed to have converged or, more accurately, when the sampler has sufficiently explored the support of the target distribution), its output can be used to approximate the posterior probabilities $\pi(\gamma|\mathbf{y}, X)$ by empirical averages based on the Gibbs output,

$$\widehat{\pi}(\gamma|\mathbf{y},X)=\left(\frac{1}{T-T_0+1}\right)\sum_{t=T_0}^{T}\mathbb{I}_{\gamma^{(t)}=\gamma}\,,$$

where the T_0 first values are eliminated as *burn-in*. (The number T_0 is therefore the assessed number of iterations needed to "reach" convergence.) The Gibbs output can also approximate the inclusion of a given variable, $\mathbb{P}(\gamma_i=1|\mathbf{y})$, as

$$\widehat{\mathbb{P}}(\gamma_i=1|\mathbf{y},X)=\left(\frac{1}{T-T_0+1}\right)\sum_{t=T_0}^{T}\mathbb{I}_{\gamma_i^{(t)}=1}\,,$$

with the same validation.

⊞ Since we have ten explanatory variables in caterpillar, it is still possible to compute the 2^{10} probabilities $\pi(\gamma|\mathbf{y},X)$ and thus deduce the normalizing constant in (3.15) for instance. We can therefore compare these true values with those produced by the Gibbs sampler. Tables 3.9 and 3.10 show that the approximation performs extremely well, using $T=20,000$ and $T_0=10,000$ in both cases. Table 3.11 provides the estimated inclusions of the variables x_i $(i=1,\ldots,10)$ in the regression for both the first-level and the noninformative-prior modelings. As noted above, the noninformative prior is more likely to include variables than the first-level one (with this particular choice of c).

Table 3.9. Dataset caterpillar: First-level G-prior model choice with $\tilde{\beta}=0_{11}$ and $c=100$, compared with the Gibbs estimates of the top ten posterior probabilities.

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y},X)$	$\widehat{\pi}(\gamma\|\mathbf{y},X)$
0, 1, 2, 4, 5	0.2316	0.2208
0, 1, 2, 4, 5, 9	0.0374	0.0375
0, 1, 9	0.0344	0.0358
0, 1, 2, 4, 5, 10	0.0328	0.0344
0, 1, 4, 5	0.0306	0.0313
0, 1, 2, 9	0.0250	0.0268
0, 1, 2, 4, 5, 7	0.0241	0.0260
0, 1, 2, 4, 5, 8	0.0238	0.0251
0, 1, 2, 4, 5, 6	0.0237	0.0244
0, 1, 2, 3, 4, 5	0.0232	0.0224

Table 3.10. Dataset caterpillar: Noninformative G-prior model choice compared with Gibbs estimates of the top ten posterior probabilities.

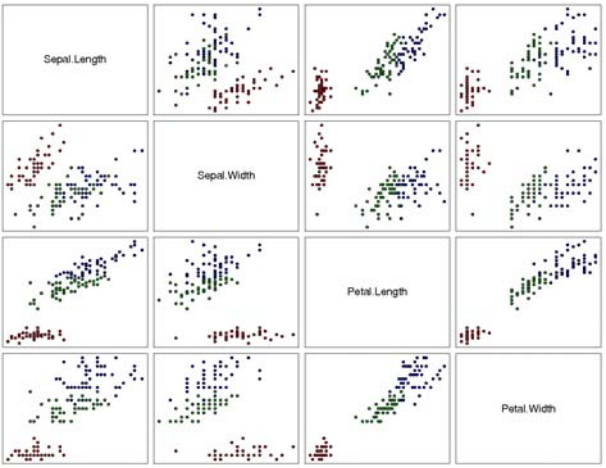
$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y},X)$	$\widehat{\pi}(\gamma\|\mathbf{y},X)$
0, 1, 2, 4, 5	0.0929	0.0929
0, 1, 2, 4, 5, 9	0.0325	0.0326
0, 1, 2, 4, 5, 10	0.0295	0.0272
0, 1, 2, 4, 5, 7	0.0231	0.0231
0, 1, 2, 4, 5, 8	0.0228	0.0229
0, 1, 2, 4, 5, 6	0.0228	0.0226
0, 1, 2, 3, 4, 5	0.0224	0.0220
0, 1, 2, 3, 4, 5, 9	0.0167	0.0182
0, 1, 2, 4, 5, 6, 9	0.0167	0.0171
0, 1, 2, 4, 5, 8, 9	0.0137	0.0130

Table 3.11. Dataset caterpillar: First-level ($\tilde{\beta} = 0_{11}$ and $c = 100$) and noninformative G-prior variable inclusion estimates (based on the same Gibbs output as Table 3.9).

γ_i	$\widehat{\mathbb{P}}_c(\gamma_i = 1\|\mathbf{y},X)$	$\widehat{\mathbb{P}}(\gamma_i = 1\|\mathbf{y},X)$
γ_1	0.8624	0.8844
γ_2	0.7060	0.7716
γ_3	0.1482	0.2978
γ_4	0.6671	0.7261
γ_5	0.6515	0.7006
γ_6	0.1678	0.3115
γ_7	0.1371	0.2880
γ_8	0.1555	0.2876
γ_9	0.4039	0.5168
γ_{10}	0.1151	0.2609

4

Generalized Linear Models



This was the sort of thing that impressed
Rebus: not nature, but ingenuity.
—**Ian Rankin**, ***A Question of Blood***.—

Roadmap

Generalized linear models are extensions of the linear regression model described in the previous chapter. In particular, they avoid the selection of a single transformation of the data that must achieve the possibly conflicting goals of normality and linearity imposed by the linear regression model, which is for instance impossible for binary or count responses. The trick that allows both a feasible processing and an extension of linear regression is first to turn the covariates into a real number by a linear projection and then to transform this value so that it fits the support of the response. We focus here on the Bayesian analysis of probit and logit models for binary data and log-linear models for contingency tables.

On the methodological side, we present a general MCMC method, the Metropolis–Hastings algorithm, which is used for the simulation of complex distributions where both regular and Gibbs sampling fail. This includes in particular the random walk Metropolis–Hastings algorithm, which acts like a plain vanilla MCMC algorithm.

4.1 A Generalization of the Linear Model

4.1.1 Motivation

In the previous chapter, we modeled the connection between a response variable y and a vector x of explanatory variables by a linear dependence relation with normal perturbations. There are many instances where both the linearity and the normality assumptions are not appropriate, especially when the support of y is restricted to $\mathbb{R}_+$ or $\mathbb{N}$. For instance, in dichotomous models, y takes its values in $\{0,1\}$ as it represents the indicator of occurrence of a particular event (death in a medical study, unemployment in a socioeconomic study, migration in a capture–recapture study, etc.); in this case, a linear conditional expectation $\mathbb{E}[y|\mathbf{x},\beta] = \mathbf{x}^\mathsf{T}\beta$ would be fairly cumbersome to handle, both in terms of the constraints on β and the corresponding distribution of the error $\varepsilon = y - \mathbb{E}[y|\mathbf{x},\beta]$.

⊞ The **bank** dataset we analyze in the first part of this chapter comes from Flury and Riedwyl (1988) and is made of four measurements on 100 genuine Swiss banknotes and 100 counterfeit ones. The response variable y is thus the status of the banknote, where 0 stands for genuine and 1 stands for counterfeit, while the explanatory factors are the length of the bill x_1, the width of the left edge x_2, the width of the right edge x_3, and the bottom margin width x_4, all expressed in millimeters. We want a probabilistic model that predicts the type of banknote (i.e., that detects counterfeit banknotes) based on the four measurements above.

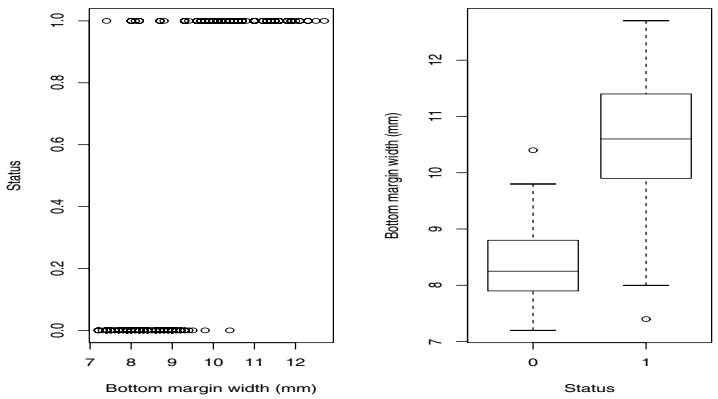


Fig. 4.1. Dataset bank: (left) Plot of the status indicator versus the bottom margin width; (right) boxplots of the bottom margin width for both counterfeit statuses.

To motivate the introduction of the generalized linear models, we only consider here the dependence of y on the fourth measure, x_4, which again is the bottom margin

width of the banknote. To start, the y_i's being binary, the conditional distribution of y given x_4 cannot be normal. Nonetheless, as shown by Figure 4.1, the variable x_4 clearly has a strong influence on whether the banknote is or is not counterfeit. To model this dependence in a proper manner, we must devise a realistic (if not real!) connection between y and x_4. The fact that y is binary implies a specific form of dependence: Indeed, both its marginal and conditional distributions necessarily are Bernoulli distributions. This means that, for instance, the conditional distribution of y given x_4 is a Bernoulli $\mathscr{B}(p(x_4))$ distribution; that is, for $x_4 = x_{4i}$, there exists $0 \leq p_i = p(x_{4i}) \leq 1$ such that

$$\mathbb{P}(y_i = 1 | x_4 = x_{4i}) = p_i \,,$$

which turns out to be also the conditional expectation of y_i, $\mathbb{E}[y_i|x_{4i}]$. If we do impose a linear dependence on the p_i's, namely,

$$p(x_{4i}) = \beta_0 + \beta_1 x_{4i} \,,$$

the maximum likelihood estimates of β_0 and β_1 are then equal to -2.02 and 0.268, leading to the estimated prediction equation

$$\hat{p}_i = -2.02 + 0.268 x_{i4} \,. \tag{4.1}$$

This implies that a banknote with bottom margin width equal to 8 is counterfeit with probability

$$-2.02 + 0.268 \times 8 = 0.120 \,.$$

Thus, this banknote has a relatively small probability of having been counterfeited, which coincides with the intuition drawn from Figure 4.1. However, if we now consider a banknote with bottom margin width equal to 12, (4.1) implies that this banknote is counterfeited with probability

$$-2.02 + 0.268 \times 12 = 1.192 \,,$$

which is certainly embarrassing for a probability estimate! We could try to modify the result by truncating the probability to $(0, 1)$ and by deciding that this value of x_4 almost certainly indicates a counterfeit, but still there is a fundamental difficulty with this model. The fact that an ordinary linear dependence can predict values outside $(0, 1)$ suggests that the connection between this explanatory variable and the probability of a counterfeit cannot be modeled through a linear function but rather can be achieved using functions of x_{4i} that take their values within the interval $(0, 1)$.

Exercise 4.1. For bank, derive the maximum likelihood estimates of β_0 and β_1 found in the previous analysis. Using Jeffreys prior on the parameters $(\beta_0, \beta_1, \sigma^2)$ of the linear regression model, compute the corresponding posterior expectation of (β_0, β_1).

4.1.2 Link Functions

As shown by the previous analysis, while linear models are nice to work with, they also have strong limitations. Therefore, we need a broader class of models to cover various dependence structures. The class selected for this chapter is called the family of *generalized linear models* (GLM), which has been formalized in McCullagh and Nelder (1989). This nomenclature stems from the fact that the dependence of y on $\mathbf{x}$ is partly *linear* in the sense that the conditional distribution of y given x is defined in terms of a linear combination $\mathbf{x}^\mathsf{T}\beta$ of the components of $\mathbf{x}$,

$$y|x, \beta \sim f(y|\mathbf{x}^\mathsf{T}\beta)\,.$$

As in the previous chapter, we use the notation $\mathbf{y} = (y_1, \ldots, y_n)$ for a sample of n responses and

$$X = [\mathbf{x}_1 \quad \ldots \quad \mathbf{x}_k] = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

for the $n \times k$ matrix of corresponding explanatory variables, possibly with $x_{11} = \ldots = x_{n1} = 1$. We use y and $\mathbf{x}$ as generic notations for single-response and covariate vectors, respectively.

A *generalized linear model* is specified by two functions:

(i) a conditional density f of y given $\mathbf{x}$ that belongs to an exponential family (Section 2.2.2) and that is parameterized by an expectation parameter $\mu = \mu(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ and possibly a dispersion parameter $\varphi > 0$ that does not depend on $\mathbf{x}$; and

(ii) a *link* function g that relates the mean $\mu = \mu(\mathbf{x})$ of f and the covariate vector, $\mathbf{x}$, as $g(\mu) = (\mathbf{x}^\mathsf{T}\beta)$, $\beta \in \mathbb{R}^k$.

For identifiability reasons, the link function g is a one-to-one function and we have

$$\mathbb{E}[y|\mathbf{x}, \beta, \varphi] = g^{-1}\left(\mathbf{x}^\mathsf{T}\beta\right)\,.$$

We can thus write the (conditional) likelihood as

$$\ell(\beta, \varphi|\mathbf{y}, X) = \prod_{i=1}^{n} f\left(y_i|\mathbf{x}^{i\mathsf{T}}\beta, \varphi\right)$$

if we choose to reparameterize f with the transform $g(\mu_i)$ of its mean and if we denote by $\mathbf{x}^i$ the covariate vector for the ith observation.[1]

[1] This upper indexing allows for the distinction between x_i, the ith component of the covariate vector, and $\mathbf{x}^i$, the ith vector of covariates in the sample.

The ordinary linear regression is obviously a special case of a GLM where $g(x) = x$, $\varphi = \sigma^2$ and, $y|x, \beta, \sigma^2 \sim \mathcal{N}\left(x^{\mathsf{T}}\beta, \sigma^2\right)$. However, outside the linear model, the interpretation of the coefficients β_i is much more difficult because they do not relate directly to the observables due to the presence of a link function that cannot be the identity. For instance, in logistic regression (defined in Example 4.1), the linear dependence is defined in terms of the *log-odds ratio* $\log\{p_1/(1-p_i)\}$.

Example 4.1. The most widely used GLMs are presumably those that analyze binary data, as in bank, that is, when $y_i|\mathbf{x}^i \sim \mathscr{B}(1, p_i)$ (with $\mu_i = p_i = p(\mathbf{x}^{i\mathsf{T}}\beta)$). The mean function p thus transforms a real value into a value between 0 and 1, and a possible choice of link function is the *logit transform*,

$$g(p) = \log(p/(1-p))\,,$$

associated with the *logistic regression model*. Because of the limited support of the responses y_i, there is no dispersion parameter in this model and the corresponding likelihood function is

$$\begin{aligned}\ell(\beta|\mathbf{y}, X) &= \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}^{i\mathsf{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathsf{T}}\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(\mathbf{x}^{i\mathsf{T}}\beta)}\right)^{1-y_i} \\ &= \exp\left\{\sum_{i=1}^{n} y_i\, \mathbf{x}^{i\mathsf{T}}\beta\right\} \Big/ \prod_{i=1}^{n} \left[1+\exp(\mathbf{x}^{i\mathsf{T}}\beta)\right]\,.\end{aligned}$$

It thus fails to factorize conveniently because of the denominator: There is no sufficient statistic of fixed dimension and therefore no manageable conjugate prior for this model, called the *logit model*. ◀

Exercise 4.2. Show that, in the setting of Example 4.1, the statistic $\sum_{i=1}^{n} y_i\, \mathbf{x}^i$ is sufficient when conditioning on the $\mathbf{x}^i$'s $(1 \le i \le n)$, and give the corresponding family of conjugate priors.

There exists a specific form of link function for each exponential family which is called the *canonical link*. This canonical function is chosen as the function $g^\star$ of the expectation parameter that appears in the exponent of the natural exponential family representation of the probability density, namely

$$g^\star(\mu) = \theta \quad \text{if} \quad f(y|\mu, \varphi) = h(y) \exp \varphi\{T(y)\cdot\theta - \Psi(\theta)\}\,.$$

Since the logistic regression model can be written as

$$f(y_i|p_i) = \exp\left\{y_i\, \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right\}\,,$$

the logit link function is the canonical version for the Bernoulli model. Note that, while it is customary to use the canonical link, there is no compelling reason to do so, besides following custom!

Example 4.2. (Example 4.1 continued) For binary response variables, many link functions can be substituted for the logit link function. For instance, the *probit* link function, $g(\mu_i) = \Phi^{-1}(\mu_i)$, where Φ is the standard normal cdf, is often used in econometrics. The corresponding likelihood is

$$\ell(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathsf{T}}\beta)^{y_i} \left[1 - \Phi(\mathbf{x}^{i\mathsf{T}}\beta)\right]^{1-y_i} . \tag{4.2}$$

Although this alternative is also quite arbitrary and any other cdf could be used as a link function (such as the logistic cdf in Example 4.1), the probit link function enjoys a missing-data (Chapter 6) interpretation that boosted its popularity: It can indeed be interpreted as a degraded linear regression model in the sense that $y_i = 1$ corresponds to the case $z_i \geq 0$, where z_i is a latent (that is, unobserved) variable such that $z_i \sim \mathcal{N}\left(\mathbf{x}^{i\mathsf{T}}\beta, 1\right)$. In other words, y appears as a dichotomized linear regression response. Of course, this perspective is only an *interpretation* of the probit model in the sense that there may be no hidden z_i's at all in the real world! In addition, the probit and logistic regression models have quite similar behaviors, differing mostly in the tails. ◀

Example 4.3. The *Poisson regression model* starts from the assumption that the y_i's are Poisson $\mathcal{P}(\mu_i)$ and it selects a link function connecting $\mathbb{R}^+$ bijectively with $\mathbb{R}$, such as, for instance, the logarithmic function, $g(\mu_i) = \log(\mu_i)$. This model is thus a *count* model in the sense that the responses are integers, for instance the number of deaths due to lung cancer in a county or the number of speeding tickets issued on a particular stretch of highway, and it is quite common in epidemiology. The corresponding likelihood is

$$\ell(\beta|\mathbf{y}, X) = \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \exp\left\{y_i\, \mathbf{x}^{i\mathsf{T}}\beta - \exp(\mathbf{x}^{i\mathsf{T}}\beta)\right\} ,$$

where the factorial terms $(1/y_i!)$ are irrelevant for both likelihood and posterior computations. (Note that it does not factorize conveniently because of the exponential terms within the exponential.) ◀

Exercise 4.3. Show that the logarithmic link is the canonical link function in the case of the Poisson regression model.

Exercise 4.4. Suppose $y_1, \ldots, y_k$ are independent Poisson $\mathcal{P}(\mu_i)$ random variables. Show that, conditional on $n = \sum_{i=1}^{k} y_i$,

$$\mathbf{y} = (y_1, \ldots, y_k) \sim \mathcal{M}_k(n; \alpha_1, \ldots, \alpha_k) ,$$

and determine the α_i's.

These three examples are simply indications of the versatility of generalized linear modeling. In this chapter, we only discuss two types of data for which generalized linear modeling is appropriate. We refer the reader to McCullagh and Nelder (1989) and Gelman et al. (2001) for detailed coverage.

4.2 Metropolis–Hastings Algorithms

As partly shown by the previous examples, posterior inference in GLMs is much harder than in linear models, which explains the longevity and versatility of linear model studies over the past centuries! Working with a GLM typically requires specific numerical or simulation tools. We take the opportunity of this requirement to introduce a universal MCMC method called the *Metropolis–Hastings* algorithm. Its range of applicability is incredibly broad (meaning that it is by no means restricted to GLM applications) and its inclusion in the Bayesian toolbox in the early 1990s has led to considerable extensions of the Bayesian field.[2]

4.2.1 Definition

When compared with the Gibbs sampler, Metropolis–Hastings algorithms are generic (or off-the-shelf) MCMC algorithms in the sense that they can be tuned toward a much wider range of possibilities. Those algorithms are also a natural extension of standard simulation algorithms such as accept–reject (see Chapter 5) or sampling importance resampling methods since they are all based on a *proposal* distribution. However, a major difference is that, for the Metropolis–Hastings algorithms, the proposal distribution is *Markov,* with kernel density $q(x, y)$. If the *target* distribution has density π, the Metropolis–Hastings algorithm is as follows:

Algorithm 4.1. Generic Metropolis–Hastings Sampler

Initialization: Choose an arbitrary starting value $x^{(0)}$.

Iteration t $(t \geq 1)$:

1. Given $x^{(t-1)}$, generate $\tilde{x} \sim q(x^{(t-1)}, x)$.
2. Compute
$$\rho(x^{(t-1)}, \tilde{x}) = \min\left(\frac{\pi(\tilde{x})/q(x^{(t-1)}, \tilde{x})}{\pi(x^{(t-1)})/q(\tilde{x}, x^{(t-1)})}, 1\right).$$
3. With probability $\rho(x^{(t-1)}, \tilde{x})$, accept $\tilde{x}$ and set $x^{(t)} = \tilde{x}$; otherwise reject $\tilde{x}$ and set $x^{(t)} = x^{(t-1)}$.

[2] This algorithm had been used by particle physicists, including Metropolis, since the late 1940s, but, as is often the case, the connection with statistics was not made until much later!

The distribution q is also called the *instrumental* distribution. As in the accept–reject method (Section 5.4), we only need to know either π or q up to a proportionality constant since both constants cancel in the calculation of ρ. Note also the advantage of this approach compared with the Gibbs sampler: it is not necessary to use the conditional distributions of π.

The strong appeal of this algorithm is that it is rather universal in its formulation as well as in its use. Indeed, we only need to simulate from a proposal q that can be chosen quite freely. There is, however, a theoretical constraint, namely that the chain produced by this algorithm must be able to explore the support of $\pi(y)$ in a finite number of steps. As discussed below, there also are many practical difficulties that are such that the algorithm may lose its universal feature and that it may require some specific tuning for each new application.

The theoretical validation of this algorithm is the same as with other MCMC algorithms: The target distribution π is the limiting distribution of the Markov chain produced by Algorithm 4.1. This is due to the choice of the acceptance probability $\rho(x, y)$ since the so-called *detailed balance equation*

$$\pi(x)q(x, y)\rho(x, y) = \pi(y)q(y, x)\rho(y, x)$$

holds and thus implies that π is stationary by integrating out x.

Exercise 4.5. Show that the detailed balance equation also holds for the *Boltzmann* acceptance probability

$$\rho(x, y) = \frac{\pi(y)q(y, x)}{\pi(y)q(y, x) + \pi(x)q(x, y)} .$$

Ⓒ While theoretical guarantees that the algorithm converges are very high, the choice of q remains essential in practice. Poor choices of q may indeed result either in a very high rejection rate, meaning that the Markov chain $(x^{(t)})_t$ hardly moves, or in a myopic exploration of the support of π, that is, in a dependence on the starting value $x^{(0)}$ such that the chain is stuck in a neighborhood region of $x^{(0)}$. A particular choice of proposal q may thus work well for one target density but be extremely poor for another one. While the algorithm is indeed universal, it is impossible to prescribe application-independent strategies for choosing q.

We thus consider below two specific cases of proposals and briefly discuss their pros and cons (see Robert and Casella, 2004, Chapter 7, for a detailed discussion).

4.2.2 The Independence Sampler

The choice of q closest to the accept–reject method (see Algorithm 5.2) is to pick a constant q that is independent of its first argument,

$$q(x, y) = q(y)\,.$$

In that case, ρ simplifies into

$$\rho(x, y) = \min\left(1, \frac{\pi(y)/q(y)}{\pi(x)/q(x)}\right)\,.$$

In the special case in which q is proportional to π, we obtain $\rho(x, y) = 1$ and the algorithm reduces, as expected, to iid sampling from π. The analogy with the accept–reject algorithm is that the maximum of the ratio π/q is replaced with the current value $\pi(x^{(t-1)})/q(x^{(t-1)})$ but the sequence of accepted $x^{(t)}$'s is not iid because of the acceptance step.

The convergence properties of the algorithm depend on the density q. First, q needs to be positive everywhere on the support of π. Second, for good exploration of this support, it appears that the ratio π/q needs to be bounded (see Robert and Casella, 2004, Theorem 7.8). Otherwise, the chain may take too long to reach some regions with low q/π values. This constraint obviously reduces the appeal of using an independence sampler, even though the fact that it does not require an explicit upper bound on π/q may sometimes be a plus.

Exercise 4.6. For π the density of an inverse normal distribution with parameters $\theta_1 = 3/2$ and $\theta_2 = 2$,

$$\pi(x) \propto x^{-3/2} \exp(-3/2x - 2/x)\mathbb{I}_{x>0},$$

write down and implement an independence MH sampler with a Gamma proposal with parameters $(\alpha, \beta) = (4/3, 1)$ and $(\alpha, \beta) = (0.5\sqrt{4/3}, 0.5)$.

Exercise 4.7. Estimate the mean of a $\mathcal{G}a(4.3, 6.2)$ random variable using

1. direct sampling from the distribution via the R command
   `> x=rgamma(n,4.3,scale=6.2)`
2. Metropolis–Hastings with a $\mathcal{G}a(4, 7)$ proposal distribution;
3. Metropolis–Hastings with a $\mathcal{G}a(5, 6)$ proposal distribution.

In each case, monitor the convergence of the cumulated average.

This type of MH sampler is thus very model-dependent, and it suffers from the same drawbacks as the importance sampling methodology, namely that tuning the "right" proposal becomes much harder as the dimension increases.

4.2.3 The Random Walk Sampler

Since the independence sampler requires too much global information about the target distribution that is difficult to come by in complex or high-

dimensional problems, an alternative is to opt for a local gathering of information, clutching to the hope that the accumulated information will provide, in the end, the global picture. Practically, this means exploring the neighborhood of the current value $x^{(t)}$ in search of other points of interest. The simplest exploration device is based on random walk dynamics.

A *random walk* proposal is based on a symmetric transition kernel $q(x, y) = q_{RW}(y - x)$ with $q_{RW}(x) = q_{RW}(-x)$. Symmetry implies that the acceptance probability $\rho(x, y)$ reduces to the simpler form

$$\rho(x, y) = \min\left(1, \pi(y)/\pi(x)\right) .$$

The appeal of this scheme is obvious when looking at the acceptance probability, since it only depends on the target π and since this version accepts all proposed moves that increase the value of π. There is considerable flexibility in the choice of the distribution q_{RW}, at least in terms of scale (i.e., the size of the neighborhood of the current value) and tails. Note that while from a probabilistic point of view random walks usually have no stationary distribution, the algorithm biases the random walk by moving toward modes of π more often than moving away from them.

Exercise 4.8. Consider x_1, x_2, and x_3 iid $\mathscr{C}(\theta, 1)$, and $\pi(\theta) \propto \exp(-\theta^2/100)$. Show that the posterior distribution of θ, $\pi(\theta|x_1, x_2, x_3)$, is proportional to

$$\exp(-\theta^2/100)[(1 + (\theta - x_1)^2)(1 + (\theta - x_2)^2)(1 + (\theta - x_3)^2)]^{-1}$$

and that it is trimodal when $x_1 = 0$, $x_2 = 5$, and $x_3 = 9$. Using a random walk based on the Cauchy distribution $\mathscr{C}(0, \sigma^2)$, estimate the posterior mean of θ using different values of σ^2. In each case, monitor the convergence.

The ambivalence of MCMC methods like the Metropolis–Hastings algorithm is that they can be applied to virtually any target. This is a terrific plus in that they can tackle new models, but there is also a genuine danger that they simultaneously fail to converge and fail to signal that they have failed to converge! Indeed, these algorithms can produce seemingly reasonable results, with all outer aspects of stability, while they are missing major modes of the target distribution. For instance, particular attention must be paid to models where the number of parameters exceeds by far the size of the dataset.

4.2.4 Output Analysis and Proposal Design

An important problem with the implementation of an MCMC algorithm is to gauge when convergence has been achieved; that is, to assess at what point the distribution of the chain is sufficiently close to its asymptotic distribution for all practical purposes or, more practically, when it has covered the whole

support of the target distribution with sufficient regularity. The number of iterations T_0 that is required to achieve this goal is called the *burn-in* period. It is usually sensible to discard simulated values within this burn-in period in the Monte Carlo estimation so that the bias caused by the starting value is reduced. However, and this is particularly true in high dimensions, the empirical assessment of MCMC convergence is extremely delicate, to the point that it is rarely possible to be certain that an algorithm has converged.[3] Nevertheless, some partial-convergence diagnostic procedures can be found in the literature (see Robert and Casella, 2004, Chapter 12).

© A first way to assess whether or not a chain is in its stationary regime is to visually compare trace plots of sequences started at different values, as it may expose difficulties related, for instance, to multimodality. In practice, when chains of length T from two starting values have visited substantially different parts of the state space, the burn-in period for at least one of the chains should be greater than T. Note, however, that the problem of obtaining overdispersed starting values can be difficult when little is known about the target density, especially in large dimensions.

Autocorrelation plots of particular components provide in addition good indications of the chain's mixing behavior. If ρ_k ($k \in \mathbb{N}^*$) denotes the kth-order autocorrelation,

$$\rho_k = \text{cov}\left(x^{(t)}, x^{(t+k)}\right) ,$$

these quantities can be estimated from the observed chain itself,[4] at least for small values of k, and an *effective sample size* factor can be deduced from these estimates,

$$T^{\text{ess}} = T \left(1 + 2 \sum_{k=1}^{T_0} \hat{\rho}_k\right)^{-1/2} ,$$

where $\hat{\rho}_k$ is the empirical autocorrelation function. This quantity represents the sample size of an equivalent iid sample when running T iterations. Conversely, the ratio T/T^{ess} indicates the multiplying factor on the minimum number of iid iterations required to run a simulation. Note, however, that this is only a partial indicator: Chains that remain stuck in one of the modes of the target distribution may well have a high effective ratio.

While we cannot discuss at length the selection of the proposal distribution (see Robert and Casella, 2004, Chapter 7), we stress that this is an important choice that has deep consequences for the convergence properties of the simulated Markov chain and thus for the exploration of the target distribution. As for prior distributions, we advise the simultaneous use of different kernels to assess their performances on the run. When considering a random

[3] Guaranteed convergence as in accept–reject algorithms is sometimes achievable with MCMC methods using techniques such as *perfect sampling* or *renewal*. But such techniques require a much more advanced study of the target distribution and the transition kernel of the algorithm. These conditions are not met very often in practice (see Robert and Casella, 2004, Chapter 13).

[4] In R, this estimation can be conducted using the acf function.

walk proposal, for instance, a quantity that needs to be calibrated against the target distribution is the scale of this random walk. Indeed, if the variance of the proposal is too small with respect to the target distribution, the exploration of the target support will be small and may fail in more severe cases. Similarly, if the variance is too large, this means that the proposal will most often generate values that are outside the support of the target and that the algorithm will reject a large portion of attempted transitions.

It seems reasonable to tune the proposal distribution in terms of its past performances, for instance by increasing the variance if the acceptance rate is high or decreasing it otherwise (or moving the location parameter toward the mean estimated over the past iterations). This must not be implemented outside a burn-in step, though, because a permanent modification of the proposal distribution amounts to taking into account the whole past of the sequence and thus it cancels both its Markovian nature and its convergence guarantees.

Example 4.4. Consider, for illustration purposes, the standard normal distribution $\mathcal{N}(0,1)$ as a target. If we use Algorithm 4.1 with a normal random walk, i.e.,

$$\tilde{x}|x^{(t-1)} \sim \mathcal{N}\left(x^{(t-1)}, \sigma^2\right) ,$$

the performance of the sampler depends on the value of σ. Picking σ^2 equal to either 10^{-4} or 10^3, for instance, provides two extreme cases: As shown in Figure 4.2, the chain has a high acceptance rate but a low exploration ability and a high autocorrelation in the former case, while its acceptance rate is low but its ability to move around the normal range is high in the latter case (with a quickly decreasing autocorrelation). Both cases use the wrong scale, though, in that the histograms of the output are quite far from the target distribution after $10,000$ iterations, and this indicates that a much larger number of iterations must be used. A comparison with Figure 4.3, which corresponds to $\sigma = 1$, clearly makes this point but also illustrates the fact that the large variance still induces large autocorrelations. ◀

Several MCMC algorithms can be mixed together within a single algorithm using either a circular or a random design. While this construction is often suboptimal (in that the inefficient algorithms in the mixture are still used on a regular basis), it almost always brings an improvement compared with its individual components. A special case where a mixed scenario is used is the *Metropolis-within-Gibbs* algorithm: When building a Gibbs sampler, it may happen that it is difficult or impossible to simulate from some of the conditional distributions. In that case, a single Metropolis step associated with this conditional distribution (as its target) can be used instead.[5]

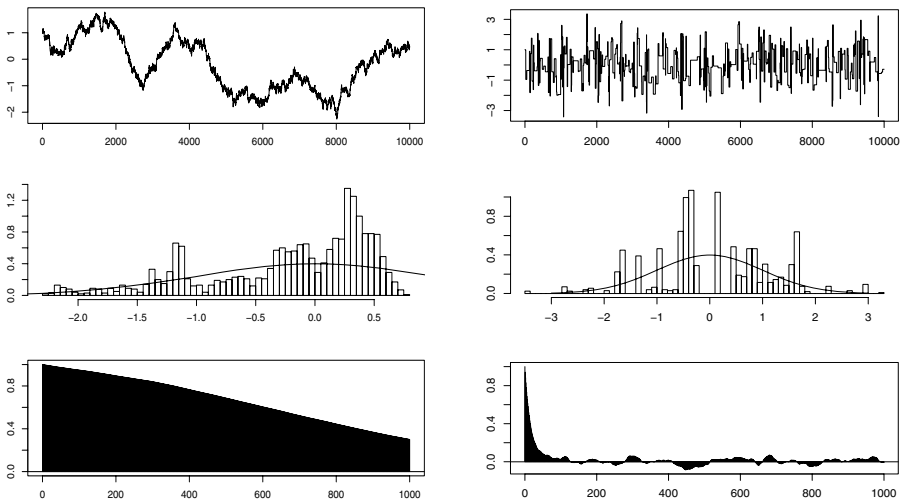


Fig. 4.2. Simulation of a $\mathcal{N}(0,1)$ target with (*left*) a $\mathcal{N}(x, 10^{-4})$ and (*right*) a $\mathcal{N}(x, 10^3)$ random walk proposal. *Top*: Sequence of 10,000 iterations subsampled at every tenth iteration; *middle*: Histogram of the last 2000 iterations compared with the target density; *bottom*: Empirical autocorrelations using R function plot.acf.

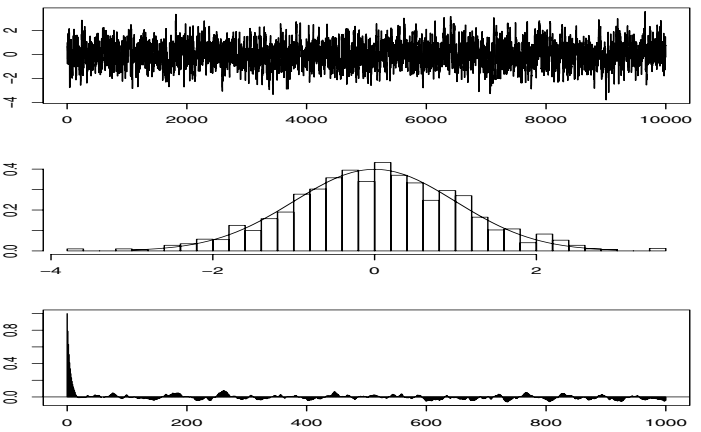


Fig. 4.3. Same legend as Figure 4.2 for a $\mathcal{N}(x, 1)$ random walk proposal.

Exercise 4.9. Rerun the experiment of Example 4.4 using instead a mixture of five random walks with variances $\sigma = 0.01, 0.1, 1, 10, 100$ and equal weights, and compare its output with the output of Figure 4.3.

4.3 The Probit Model

We now engage in a full discussion of the Bayesian processing of the probit model of Example 4.2, taking special care to differentiate between the various prior modelings.

4.3.1 Flat Prior

If no prior information is available, we can resort (as usual!) to a flat prior on β, $\pi(\beta) \propto 1$, and then obtain the posterior distribution

$$\pi(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathsf{T}}\beta)^{y_i} \left[1 - \Phi(\mathbf{x}^{i\mathsf{T}}\beta)\right]^{1-y_i} ,$$

which is nonstandard and must be simulated using MCMC techniques.

Exercise 4.10. Find conditions on the observed pairs (x_i, y_i) for the posterior distribution above to be proper.

A variety of Metropolis–Hastings algorithms have been proposed for obtaining samples from this posterior distribution. Here we consider a sampler that appears to work well when the number of predictors is reasonably small. This Metropolis–Hastings sampler is a random walk scheme that uses the maximum likelihood estimate $\hat{\beta}$ as a starting value and the asymptotic (Fisher) covariance matrix $\hat{\Sigma}$ of the maximum likelihood estimate as the covariance matrix for the proposal[6] density, $\tilde{\beta} \sim \mathcal{N}_k(\beta^{(t-1)}, \tau^2\hat{\Sigma})$.

[5] We stress that we do not resort to an MH algorithm for the purpose of simulating exactly from the corresponding conditional since this would require an infinite number of iterations but rather that we use a *single* iteration of the MH algorithm as a substitute for the simulation from the conditional since the resulting MCMC algorithm is still associated with the same stationary distribution.

[6] A choice of parameters that depend on the data for the Metropolis–Hastings proposal is entirely correct both from an MCMC point of view (meaning that this is not a self-tuning algorithm) and from a Bayesian point of view (since the parameters of the proposal are not those of the prior).

ALGORITHM 4.2. PROBIT METROPOLIS–HASTINGS SAMPLER

Initialization: Compute the MLE $\hat{\beta}$ and the covariance matrix $\hat{\Sigma}$ corresponding to the asymptotic covariance of $\hat{\beta}$, and set $\beta^{(0)} = \hat{\beta}$.

Iteration $t \geq 1$:

1. Generate $\tilde{\beta} \sim \mathcal{N}_k(\beta^{(t-1)}, \tau^2 \hat{\Sigma})$.
2. Compute
$$\rho(\beta^{(t-1)}, \tilde{\beta}) = \min\left(1, \frac{\pi(\tilde{\beta}|\mathbf{y})}{\pi(\beta^{(t-1)}|\mathbf{y})}\right).$$
3. With probability $\rho(\beta^{(t-1)}, \tilde{\beta})$, take $\beta^{(t)} = \tilde{\beta}$; otherwise take $\beta^{(t)} = \beta^{(t-1)}$.

Ⓒ The R function `glm` is very helpful for the initialization step of Algorithm 4.2. The terminology used in our R program is

```
> mod=summary(glm(y~X,family=binomial(link="probit")))
```

with `mod$coeff[,1]` containing $\hat{\beta}$ and `mod$cov.unscaled` providing $\hat{\Sigma}$.

⊞ For **bank**, using a probit modeling with no intercept over the four measurements, we tested three different scales, namely $\tau = 1, 0.1, 10$, by running Algorithm 4.2 over $10,000$ iterations. Looking both at the raw sequences and at the autocorrelation graphs, it appears that the best mixing behavior is associated with $\tau = 1$. Figure 4.4 illustrates the output of the simulation run in that case. (We did not include the graphs for the other values of τ, but the curious reader can check that there is indeed a clear difference with the case $\tau = 1$.) Using a burn-in range of 1000 iterations, the averages of the parameters over the last 9000 iterations are equal to $-1.2193, 0.9540, 0.9795$, and 1.1481, respectively. A plug-in estimate of the predictive probability of a counterfeit banknote is therefore

$$\hat{p}_i = \Phi\left(-1.2193 x_{i1} + 0.9540 x_{i2} + 0.9795 x_{i3} + 1.1481 x_{i4}\right).$$

For instance, according to this equation, a banknote of length 214.9 mm, left-edge width 130.1 mm, right-edge width 129.9 mm, and bottom margin width 9.5 mm is counterfeited with probability

$$\Phi\left(-1.1293 \times 214.9 + \ldots + 1.1481 \times 9.5\right) \approx 0.5917.$$

While the plug-in representation above gives an immediate evaluation of the predictive probability, a better approximation to this probability function is provided by the average over the iterations of the current predictive probabilities, $\Phi\left(\beta^{(t)} x_{i1} + \beta^{(t)} x_{i2} + \beta^{(t)} x_{i3} + \beta^{(t)} x_{i4}\right)$.

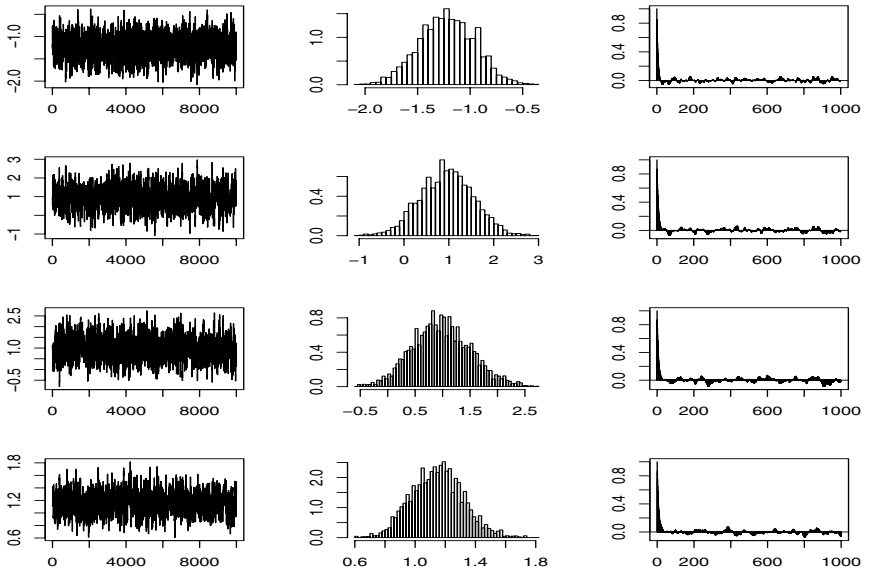


Fig. 4.4. Dataset bank: Estimation of the probit coefficients via Algorithm 4.2 and a flat prior. *Left*: β_i's $(i = 1, \ldots, 4)$; *center*: Histogram over the last 9000 iterations; *right*: Autocorrelation over the last 9000 iterations.

Exercise 4.11. Include an intercept in the probit analysis of bank and run the corresponding version of Algorithm 4.2 to discuss whether or not the posterior variance of the intercept is high.

Exercise 4.12. Using the latent variable representation of Example 4.2, introduce $z_i|\beta \sim \mathcal{N}\left(\mathbf{x}^{i\mathsf{T}}\beta, 1\right)$ $(1 \le i \le n)$ such that $y_i = \mathbb{B}_{z_i \le 0}$. Deduce that

$$z_i|y_i, \beta \sim \begin{cases} \mathcal{N}_+\left(\mathbf{x}^{i\mathsf{T}}\beta, 1, 0\right) & \text{if} \quad y_i = 1\,, \\ \mathcal{N}_-\left(\mathbf{x}^{i\mathsf{T}}\beta, 1, 0\right) & \text{if} \quad y_i = 0\,, \end{cases}$$

where $\mathcal{N}_+(\mu, 1, 0)$ and $\mathcal{N}_-(\mu, 1, 0)$ are the normal distributions with mean μ and variance 1 that are left-truncated and right-truncated at 0, respectively. Check that those distributions can be simulated using the R commands

```
> xp=qnorm(runif(1)*pnorm(mu)+pnorm(-mu))+mu
> xm=qnorm(runif(1)*pnorm(-mu))+mu
```

Under the flat prior $\pi(\beta) \propto 1$, show that

$$\beta|\mathbf{y}, \mathbf{z} \sim \mathcal{N}_k\left((X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{z}, (X^{\mathsf{T}}X)^{-1}\right)\,,$$

where $\mathbf{z} = (z_1, \ldots, z_n)$, and derive the corresponding Gibbs sampler, sometimes called the *Albert–Chib* sampler. (*Hint*: A good starting point is the maximum

likelihood estimate of β.) Compare the application to bank with the output in Figure 4.4. (*Note*: Account for differences in computing time.)

4.3.2 Noninformative *G*-Priors

Following the principles discussed in earlier chapters (see, e.g., Chapter 3), a flat prior on β is not appropriate for comparison purposes since we cannot trust the corresponding Bayes factors. In a variable selection setup, we thus replace the flat prior with a hierarchical prior,

$$\beta|\sigma^2, X \sim \mathcal{N}_k\left(0_k, \sigma^2(X^{\mathsf{T}}X)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X) \propto \sigma^{-3/2}\,,$$

inspired by the normal linear regression model, the modification in the power of σ^2 being driven by integrability constraints.[7] Integrating out σ^2 in this joint prior then leads to

$$\pi(\beta|X) \propto |X^{\mathsf{T}}X|^{1/2}\Gamma((2k-1)/4)\left(\beta^{\mathsf{T}}(X^{\mathsf{T}}X)\beta\right)^{-(2k-1)/4}\pi^{-k/2}\,,$$

which is clearly improper. Nonetheless, if we consider the *same* hierarchical prior for a submodel associated with a subset of the predictor variables in X, associated with the *same* variance factor σ^2, the marginal distribution of $\mathbf{y}$ then depends on the *same* unknown multiplicative constant as the full model, and this constant cancels in the corresponding Bayes factor. (This is exactly the same idea as for Zellner's noninformative G-prior, see Section 3.3.2.)

The corresponding posterior distribution of β is

$$\pi(\beta|\mathbf{y}, X) \propto |X^{\mathsf{T}}X|^{1/2}\Gamma((2k-1)/4)\left(\beta^{\mathsf{T}}(X^{\mathsf{T}}X)\beta\right)^{-(2k-1)/4}\pi^{-k/2} \times \prod_{i=1}^{n} \Phi(x^{i\mathsf{T}}\beta)^{y_i}\left[1-\Phi(\mathbf{x}^{i\mathsf{T}}\beta)\right]^{1-y_i}\,. \tag{4.3}$$

Note that we need to keep the terms $|X^{\mathsf{T}}X|^{1/2}$, $\Gamma((2k-1)/4)$, and $\pi^{-k/2}$, in this expression because they vary among submodels. To omit these terms would result in a bias in the computation of the Bayes factors.

Exercise 4.13. Find conditions on $\sum_i y_i$ and $\sum_i(1-y_i)$ for the posterior distribution defined by (4.3) to be proper.

Contrary to the linear regression setting and as for the flat prior in Section 4.3.1, neither the posterior distribution of β nor the marginal distribution of $\mathbf{y}$

[7] Note that the matrix $X^{\mathsf{T}}X$ is *not* the Fisher information matrix outside the normal model but that the Fisher information matrix usually involves a function of β that prevents its use as a prior (inverse) covariance matrix on β.

can be derived analytically. We can however use exactly the same Metropolis–Hastings sampler as in Section 4.3.1, namely a random walk proposal based on the estimated Fisher information matrix and the MLE $\hat{\beta}$ as the starting value.

⊞ For **bank**, the corresponding approximate Bayes estimate of β is given by

$$\mathbb{E}^{\pi}[\beta|\mathbf{y}, X] \approx (-1.1552, 0.9200, 0.9121, 1.0820),$$

which slightly differs from the estimate found in Section 4.3.1 for the flat prior. This approximation was obtained by running the MH algorithm with scale $\tau^2 = 1$ over $10,000$ iterations and averaging over the last 9000 iterations. Figure 4.5 gives an assessment of the convergence of the MH scheme that does not vary very much compared with the previous figure.

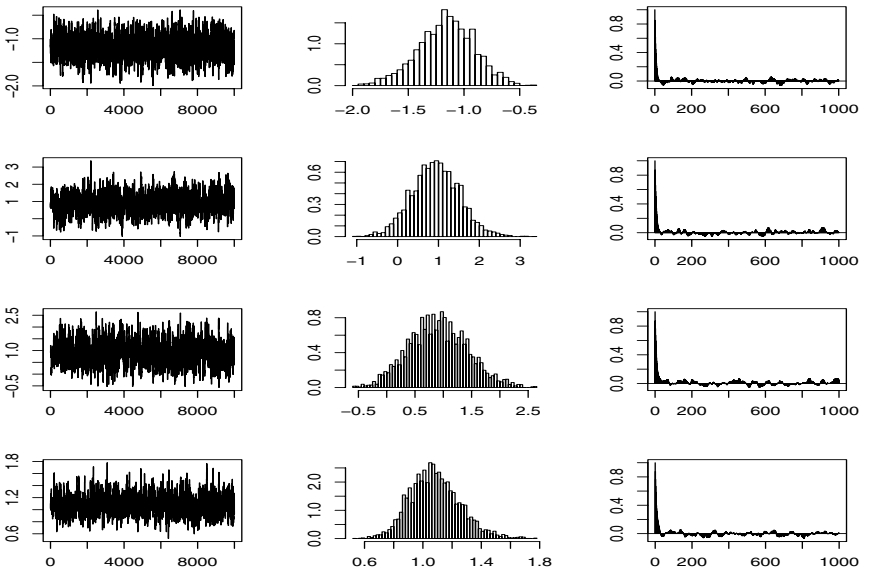


Fig. 4.5. Dataset bank: Same legend as Figure 4.4 using an MH algorithm and a G-prior on β.

We now address the specific problem of approximating the marginal distribution of $\mathbf{y}$ in order to provide approximations to the Bayes factor and thus achieve the Bayesian equivalent of standard software to identify significant variables in the probit model. The marginal distribution of $\mathbf{y}$ is

$$f(\mathbf{y}|X) \propto |X^{\mathsf{T}}X|^{1/2}\,\pi^{-k/2}\Gamma((2k-1)/4)\int \left(\beta^{\mathsf{T}}(X^{\mathsf{T}}X)\beta\right)^{-(2k-1)/4} \times \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathsf{T}}\beta)^{y_i}\left[1-\Phi(\mathbf{x}^{i\mathsf{T}}\beta)\right]^{1-y_i}\,\mathrm{d}\beta\,,$$

which cannot be computed in closed form. We thus propose to use as a generic proxy an importance sampling approximation to this integral based on a normal approximation $\mathscr{N}_k(\widehat{\beta}, 2\,\widehat{V})$ to $\pi(\beta|\mathbf{y}, X)$, where $\widehat{\beta}$ is the MCMC approximation of $\mathbb{E}^{\pi}[\beta|\mathbf{y}, X]$ and $\widehat{V}$ is the MCMC approximation[8] of $\mathbb{V}(\beta|\mathbf{y}, X)$. The corresponding estimate of the marginal distribution of $\mathbf{y}$ is then, up to a constant,

$$\frac{\Gamma(\frac{2k-1}{4})|X^{\mathsf{T}}X|^{1/2}}{\pi^{k/2}M}\sum_{m=1}^{M}\left(\beta^{(m)\mathsf{T}}(X^{\mathsf{T}}X)\beta^{(m)}\right)^{\frac{-2k+1}{4}}\prod_{i=1}^{n}\Phi(\mathbf{x}^{i\mathsf{T}}\beta^{(m)})^{y_i} \times\left[1-\Phi(\mathbf{x}^{i\mathsf{T}}\beta^{(m)})\right]^{1-y_i}|\widehat{V}|^{1/2}(4\pi)^{k/2}\,e^{(\beta^{(m)}-\widehat{\beta})^{\mathsf{T}}\widehat{V}^{-1}(\beta^{(m)}-\widehat{\beta})/4}\,, \qquad (4.4)$$

where the $\beta^{(m)}$'s are simulated from the $\mathscr{N}_k(\widehat{\beta}, 2\,\widehat{V})$ importance distribution.

If we consider a linear restriction on β such as $H_0 : R\beta = r$, with $r \in \mathbb{R}^q$ and R a $q \times k$ matrix of rank q, the submodel is associated with the likelihood

$$\ell(\beta^0|\mathbf{y}, X_0) \propto \prod_{i=1}^{n}\Phi(x_0^{i\mathsf{T}}\beta^0)^{y_i}\left[1-\Phi(x_0^{i\mathsf{T}}\beta^0)\right]^{1-y_i}\,,$$

where β^0 is $(k-q)$-dimensional and X_0 and x_0 are linear transforms of X and x of dimensions $(n, k-q)$ and $(k-q)$, respectively. Under the G-prior

$$\beta^0|\sigma^2, X_0 \sim \mathscr{N}_{k-q}\left(0_{k-q}, \sigma^2(X_0^{\mathsf{T}}X_0)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X_0) \propto \sigma^{-3/2}\,,$$

the marginal distribution of $\mathbf{y}$ is of the same type as in the unconstrained case, namely,

$$f(y|X_0) \propto |X_0^{\mathsf{T}}X_0|^{1/2}\pi^{-(k-q)/2}\Gamma\{\frac{2(k-q)-1}{4}\}\int\left\{(\beta^0)^{\mathsf{T}}(X_0^{\mathsf{T}}X_0)\beta^0\right\}^{\frac{-2(k-q)+1}{4}} \times\prod_{i=1}^{n}\Phi(x_0^{i\mathsf{T}}\beta^0)^{y_i}\left[1-\Phi(x_0^{i\mathsf{T}}\beta^0)\right]^{1-y_i}\mathrm{d}\beta^0\,.$$

Once again, if we first run an MCMC sampler for the posterior of β^0 for this submodel, it provides both parameters of a normal importance distribution and thus allows an approximation of the marginal distribution of $\mathbf{y}$ in the submodel in all ways similar to (4.4).

[8] The factor 2 in the covariance matrix allows some amount of overdispersion, which is always welcomed in importance sampling settings, if only for variance finiteness purposes.

⊞ For **bank**, if we want to test the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, we obtain $B_{10}^{\pi} = 8916.0$ via the importance sampling approximation of (4.4). Using Jeffreys' scale of evidence, since $\log_{10}(B_{10}^{\pi}) = 3.950$, the posterior distribution is strongly against H_0.

More generally, we can produce a Bayesian regression output, programmed in R, that mimics the standard software output for generalized linear models. Along with the estimates of the β_i's, given by their posterior expectation, we include the posterior variances of the β_i's, also derived from the MCMC sample, and the log Bayes factors $\log_{10}\left(B_{10}^{i}\right)$ corresponding to the null hypotheses $H_0 : \beta_i = 0$. As above, the Bayes factors are computed by importance sampling based on $100,000$ simulations. The stars are related to Jeffreys' scale of evidence.

```
             Estimate  Post. var.  log10(BF)

X1           -1.1552   0.0631       4.5844 (****)
X2            0.9200   0.3299      -0.2875
X3            0.9121   0.2595      -0.0972
X4            1.0820   0.0287      15.6765 (****)

evidence against H0: (****) decisive, (***) strong,
(**) substantial, (*) poor
```

Although these Bayes factors cannot be used simultaneously, an informal conclusion is that the important variables for the prediction of counterfeiting seem to be X_1 and X_4.

Exercise 4.14. For bank, compute the Bayes factor associated with the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$.

4.3.3 About Informative Prior Analyses

In the setting of probit (and other generalized linear) models, it is unrealistic to expect practitioners to come up with precise prior information about the parameters β. There exists nonetheless an amenable approach to prior information through what is called the *conditional mean family of prior distributions*. The intuition behind this approach is that prior beliefs about the probabilities p_i can be assessed to some extent by the practitioners for *particular values* of the explanatory variables $x_{1i}, \ldots, x_{ki}$. Once this information is taken into account, a corresponding prior can be derived for the parameter vector β. This technique is certainly one of the easiest methods of incorporating subjective prior information into the processing of the binary regression problem, especially because it appeals to practitioners for whom the β's have, at best, a virtual meaning.

Starting with k explanatory variables, we derive the subjective prior information from k different values of the covariate vector, denoted by $\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^k$.

(The theoretical motivation for setting the number of covariate vectors exactly equal to the dimension of β will be made clear below.) For each of these values, the practitioner is asked to specify two things:

(i) a prior guess g_i at the probability of success p_i associated with $\mathbf{x}^i$; and
(ii) an assessment of her or his certainty about that guess translated as a number K_i of equivalent "prior observations."[9] This question can be expressed as "On how many imaginary observations did you build this guess?"

Both quantities can be turned into a formal prior density on β by imposing a beta prior distribution on p_i with parameters $K_i g_i$ and $K_i(1-g_i)$ since the mean of a $\mathscr{B}e(a,b)$ distribution is $a/(a+b)$. If we make the additional assumption that the k probabilities $p_1,\ldots,p_k$ are a priori independent (which clearly does not hold since they all depend on the same β!), their joint density is

$$\pi(p_1,\ldots,p_k) \propto \prod_{i=1}^{k} p_i^{K_i g_i - 1}(1-p_i)^{K_i(1-g_i)-1}\,. \tag{4.5}$$

Now, if we relate the probabilities p_i to the parameter β, conditional on the covariate vectors $\tilde{\mathbf{x}}^1,\ldots,\tilde{\mathbf{x}}^k$, by $p_i = \Phi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)$, we conclude that the corresponding distribution on β is

$$\pi(\beta) \propto \prod_{i=1}^{k} \Phi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)^{K_i g_i - 1}\left[1-\Phi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)\right]^{K_i(1-g_i)-1}\varphi(\tilde{\mathbf{x}}^{i\mathsf{T}}\beta)\,.$$

(This change of variable explains why we needed exactly k different covariate vectors in the prior assessment.)

Exercise 4.15. Compute the Jacobian $|\partial p_1 \cdots \partial p_k / \partial\beta_1 \cdots \partial\beta_k|$ and deduce that the transform of the prior density $\pi(p_1,\ldots,p_k)$ in the prior density $\pi(\beta)$ above is correct.

This intuitive approach to prior modeling is also interesting from a computational point of view since the corresponding posterior distribution

$$\begin{aligned}\pi(\beta|\mathbf{y},X) \propto &\prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathsf{T}}\beta)^{y_i}\left[1-\Phi(\mathbf{x}^{i\mathsf{T}}\beta)\right]^{1-y_i}\\ &\times\prod_{j=1}^{k} \Phi(\tilde{\mathbf{x}}^{j\mathsf{T}}\beta)^{K_j g_j - 1}\left[1-\varphi(\tilde{\mathbf{x}}^{j\mathsf{T}}\beta)\right]^{K_j(1-g_j)-1}\varphi(\tilde{\mathbf{x}}^{j\mathsf{T}}\beta)\end{aligned}$$

is of almost exactly the same type as the posterior distributions in both noninformative modelings above. The main difference stands in the product of the Jacobian terms $\varphi(\tilde{\mathbf{x}}^{j\mathsf{T}}\beta)$ $(1 \le j \le k)$, but

[9] This technique is called the device of *imaginary observations* and was proposed by the Italian statistician Bruno de Finetti.

$$\prod_{j=1}^{k} \varphi(\tilde{\mathbf{x}}^{j\mathsf{T}}\beta) \propto \exp\left\{-\sum_{j=1}^{k}(\tilde{\mathbf{x}}^{j\mathsf{T}}\beta)^2/2\right\} = \exp\left\{-\beta^{\mathsf{T}}\left[\sum_{j=1}^{k}\tilde{\mathbf{x}}^{j}\tilde{\mathbf{x}}^{j\mathsf{T}}\right]\beta/2\right\}$$

means that, if we forget about the -1's in the exponents, this posterior distribution corresponds to a regular posterior distribution for the probit model when adding to the observations $(y_1, \mathbf{x}^1), \ldots, (y_n, \mathbf{x}^n)$ the pseudo-observations[10] $(g_1, \tilde{\mathbf{x}}^1), \ldots, (g_1, \tilde{\mathbf{x}}^1), \ldots, (g_k, \tilde{\mathbf{x}}^k), \ldots, (g_k, \tilde{\mathbf{x}}^k)$, where each pair $(g_i, \tilde{\mathbf{x}}^i)$ is repeated K_i times and when using the G-prior

$$\beta \sim \mathcal{N}_k\left(0_k, \left[\sum_{j=1}^{k}\tilde{\mathbf{x}}^{j}\tilde{\mathbf{x}}^{j\mathsf{T}}\right]^{-1}\right).$$

Therefore, Algorithm 4.2 does not need to be adapted to this case.

4.4 The Logit Model

We now briefly repeat the developments of the previous section in the case of the logit model, as defined in Example 4.1, not because there exist notable differences with either the processing or the conclusions of the probit model but rather because there is hardly any difference. For instance, Algorithm 4.2 can also be used for this setting, based on the same proposal but simply modifying the definition of $\pi(\beta|\mathbf{y})$, since the likelihood is now

$$\ell(\beta|\mathbf{y}, X) = \exp\left\{\sum_{i=1}^{n} y_i\, \mathbf{x}^{i\mathsf{T}}\beta\right\} \Big/ \prod_{i=1}^{n}\left[1 + \exp(\mathbf{x}^{i\mathsf{T}}\beta)\right]. \tag{4.6}$$

That both models can be processed in a very similar manner means, for instance, that they can be easily compared when one is uncertain about which link function to adopt. The Bayes factor used in the comparison of the probit and logit models is directly derived from the importance sampling experiments described for the probit model. Note also that, while the values of the parameter β differ between the two models, a subjective prior modeling as in Section 4.3.3 can be conducted simultaneously for both models, the only difference occurring for the change of variables from $(p_1, \ldots, p_k)$ to β.

Exercise 4.16. In the case of the logit model—i.e., when $p_i = \exp\tilde{\mathbf{x}}^{i\mathsf{T}}\beta/\{1 + \exp\tilde{\mathbf{x}}^{i\mathsf{T}}\beta\}$ $(1 \le i \le k)$—derive the prior distribution on β associated with the prior (4.5) on $(p_1, \ldots, p_k)$.

[10] Note that the fact that the g_j's do not take their values in $\{0, 1\}$ but rather in $(0, 1)$ does not create any difficulty in the implementation of Algorithm 4.2.

If we use a flat prior on β, the posterior distribution proportional to (4.6) can be inserted directly in Algorithm 4.2 to produce a sample approximately distributed from this posterior (assuming the conditions for its existence are met).

⊞ For **bank**, Figure 4.6 summarizes the results of running Algorithm 4.2 with the scale factor equal to $\tau = 1$: There is no clear difference between these graphs and those of earlier figures, except for a slight increase in the skewness of the histograms of the β_i's. (Obviously, this does not necessarily reflect a different convergence behavior but possibly a different posterior behavior since we are not dealing with the *same* posterior distribution.) The MH approximation—based on the last 9000 iterations—of the Bayes estimate of β is equal to $(-2.5888, 1.9967, 2.1260, 2.1879)$. We can note the numerical difference between these values and those produced by the probit model. The sign and the relative magnitudes of the components are, however, very similar. For comparison purposes, consider the plug-in estimate of the predictive probability of a counterfeit banknote,

$$\hat{p}_i = \frac{\exp\left(-2.5888x_{i1} + 1.9967x_{i2} + 2.1260x_{i3} + 2.1879x_{i4}\right)}{1 + \exp\left(-2.5888x_{i1} + 1.9967x_{i2} + 2.1260x_{i3} + 2.1879x_{i4}\right)}.$$

Using this approximation, a banknote of length 214.9 mm, of left-edge width 130.1 mm, of right-edge width 129.9 mm, and of bottom margin width 9.5 mm is counterfeited with probability

$$\frac{\exp\left(-2.5888 \times 130.1 + \ldots + 2.1879 \times 9.5\right)}{1 + \exp\left(-2.5888 \times 130.1 + \ldots + 2.1879 \times 9.5\right)} \approx 0.5963\,.$$

This estimate of the probability is therefore very close to the estimate derived from the probit modeling, which was equal to 0.5917 (especially if we take into account the uncertainties associated both with the MCMC experiments and with the plug-in shortcut).

For model comparison purposes and the computation of Bayes factors, we can also use the same G-prior as for the probit model and thus multiply (4.6) by $|X^{\mathsf{T}}X|^{1/2}\Gamma((2k-1)/4)\left(\beta^{\mathsf{T}}(X^{\mathsf{T}}X)\beta\right)^{-(2k-1)/4}\pi^{-k/2}$. The MH implementation obviously remains the same.

Exercise 4.17. Examine whether or not the sufficient conditions for propriety of the posterior distribution found in Exercise 4.13 for the probit model are the same for the logit model.

⊞ For **bank**, Figure 4.7 once more summarizes the output of the MH scheme over $10,000$ iterations. (Since we observe the same skewness in the histograms as in Figure 4.6, this feature is most certainly due to the corresponding posterior distribution rather than to a deficiency in the convergence of the algorithm.) We can repeat the test of the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ already done for the probit model and then obtain an approximate Bayes factor of $B^{\pi}_{10} = 16972.3$, with the same conclusion as earlier (although with twice as large an absolute value). We can also take advantage of the output software programmed for the probit model to produce the following summary:

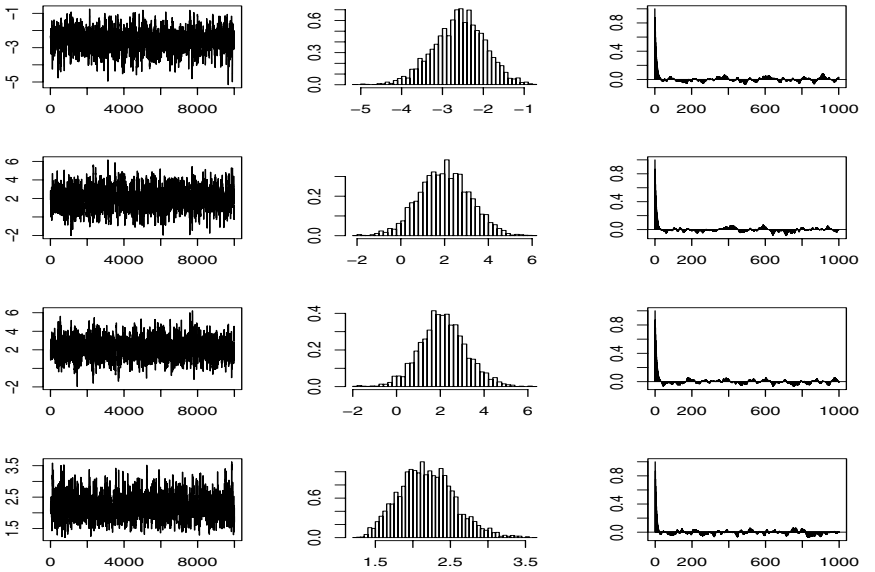


Fig. 4.6. Dataset bank: Estimation of the logit coefficients via Algorithm 4.2 under a flat prior. *Left*: β_i's ($i = 1, \ldots, 4$); *center*: Histogram over the last 9000 iterations; *right*: Autocorrelation over the last 9000 iterations.

```
            Estimate  Post. var.  log10(BF)

X1          -2.3970   0.3286       4.8084 (****)
X2           1.6978   1.2220      -0.2453
X3           2.1197   1.0094      -0.1529
X4           2.0230   0.1132      15.9530 (****)

evidence against H0: (****) decisive, (***) strong,
(**) substantial, (*) poor
```

Therefore, the most important covariates are again X_1 and X_4.

Exercise 4.18. For **bank** and the logit model, compute the Bayes factor associated with the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ and compare its value with the value obtained for the probit model in Exercise 4.14.

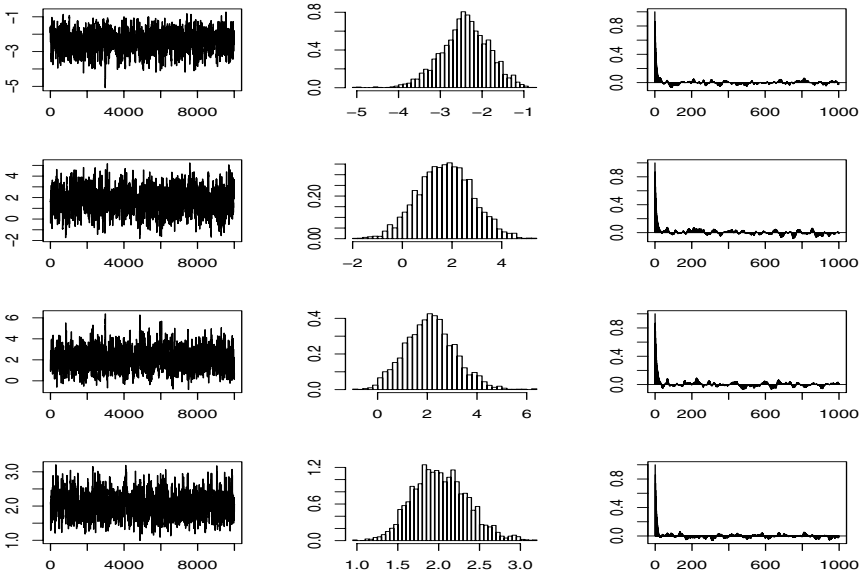


Fig. 4.7. Dataset bank: Same legend as Figure 4.6 using an MH algorithm and a G-prior on β.

4.5 Log-Linear Models

We conclude this chapter with an application of generalized linear modeling to the case of factors, already mentioned in Section 3.1.1. A standard approach to the analysis of associations (or dependencies) between *categorical* variables (that is, variables that take a finite number of values) is to use *log-linear models*. These models are special cases of generalized linear models connected to the Poisson distribution, and their name stems from the fact that they have traditionally been based on the logarithmic link function (see Example 4.3).

4.5.1 Contingency Tables

In such models, a sufficient statistic is the *contingency table*, which is a multiple-entry table made up of the cross-classified counts for the different categorical variables. There is much literature on contingency tables, including for instance Whittaker (1990) and Agresti (1996), because the corresponding models are quite handy both in the social sciences and in survey processing, where the observables are always reduced to a finite number of values.

⊞ The **airquality** dataset was obtained from the New York State Department of Conservation (ozone data) and from the American National Weather Service (meteorological

data) and is part of the datasets contained in R (Chambers et al., 1983) and available as

```
> air=data(airquality)
```

This dataset involves two repeated measurements over 111 consecutive days, namely the mean ozone u (in parts per billion) from 1 pm to 3 pm at Roosevelt Island, the maximum daily temperature v (in degrees F) at La Guardia Airport, and, in addition, the month w (coded from 5 for May to 9 for September). If we discretize the measurements u and v into dichotomous variables (using the empirical median as the cutting point), we obtain the following three-way contingency table of counts per combination of the three (discretized) factors:

```
              month  5  6  7  8  9
ozone     temp
[1,31]    [57,79] 17  4  2  5 18
          (79,97]  0  2  3  3  2
(31,168] [57,79]  6  1  0  3  1
          (79,97]  1  2 21 12  8
```

This contingency table thus has $5 \times 2 \times 2 = 20$ entries deduced from the number of categories of the three factors, among which some are zero because the corresponding combination of the three factors has not been observed in the study.

Each term in the table being an integer, it can then in principle be modeled as a Poisson variable. If we denote the counts by $\mathbf{y} = (y_1, \ldots, y_n)$, where $i = 1, \ldots, n$ is an arbitrary way of indexing the cells of the table, we can thus assume that $y_i \sim \mathscr{P}(\mu_i)$. Obviously, the likelihood

$$\ell(\mu|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\mu_i!} \mu_i^{y_i} \exp(-\mu_i)\,,$$

where $\mu = (\mu_1, \ldots, \mu_n)$, shows that the model is *saturated*, namely that no structure can be exhibited because there are as many parameters as there are observations. To exhibit any structure, we need to constrain the μ_i's and do so via a GLM whose covariate matrix X is directly derived from the contingency table itself.

If some entries are structurally equal to zero (as for instance when crossing number of pregnancies with male indicators), these entries should be removed from the model.

Exercise 4.19. In the case of a 2×2 contingency table with fixed total count $n = n_{11} + n_{12} + n_{21} + n_{22}$, we denote by $\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}$ the corresponding probabilities. If the prior on those probabilities is a Dirichlet $\mathcal{D}_4(1/2, \ldots, 1/2)$, give the corresponding marginal distributions of $\alpha = \theta_{11} + \theta_{12}$ and $\beta = \theta_{11} +$

θ_{21}. Deduce the associated Bayes factor if H_0 is the hypothesis of independence between the factors and if the priors on the margin probabilities α and β are those derived above.

When we express the mean parameters μ_i of a log-linear model as

$$\log(\mu_i) = \mathbf{x}^{i\mathsf{T}}\beta\,,$$

the covariate vector $\mathbf{x}^i$ is indeed quite special in that it is made up *only* of indicators. The so-called *incidence matrix* X with rows equal to the $\mathbf{x}^i$'s is thus such that its elements are all zeros or ones. Given a contingency table, the choice of indicator variables to include in $\mathbf{x}^i$ can vary, depending on what is deemed (or found) to be an important relation between some categorical variables. For instance, suppose that there are three categorical variables, u, v, and w as in airquality, and that u takes I values, v takes J values, and w takes K values. If we only include the indicators for the values of the three categorical variables in X, we have

$$\log(\mu_\tau) = \sum_{b=1}^{I} \beta_b^u \mathbb{I}_b(u_\tau) + \sum_{b=1}^{J} \beta_b^v \mathbb{I}_b(v_\tau) + \sum_{b=1}^{K} \beta_b^w \mathbb{I}_b(w_\tau)\,;$$

that is, $(1 \le i \le I,\, 1 \le j \le J,\, 1 \le k \le K)$,

$$\log(\mu_{l(i,j,k)}) = \beta_i^u + \beta_j^v + \beta_k^w$$

$(1 \le i \le I,\, 1 \le j \le J,\, 1 \le k \le K)$, where $l(i,j,k)$ corresponds to the index of the (i,j,k) entry in the table, namely the case when $u = i$, $v = j$, and $w = k$. Similarly, the saturated log-linear model corresponds to the use of one indicator per entry of the table; that is,

$$\log(\mu_{l(i,j,k)}) = \beta_{ijk}^{uvw}$$

$(1 \le i \le I,\, 1 \le j \le J,\, 1 \le k \le K)$.

For comparison reasons that will very soon be apparent, and by analogy with analysis of variance (ANOVA) conventions, we can also over-parameterize this representation as

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} + \lambda_{ijk}^{uvw}\,, \tag{4.7}$$

where λ appears as the overall or reference average effect, λ_i^u appears as the marginal discrepancy (against the reference effect λ) when $u = i$, λ_{ij}^{uv} as the interaction discrepancy (against the added effects $\lambda + \lambda_i^u + \lambda_j^v$) when $(u,v) = (i,j)$, etc.

Using the representation (4.7) is quite convenient because it allows a straightforward parameterization of the nonsaturated models, which then appear as submodels of (4.7) where some groups of parameters are null. For example,

(i) if both categorical variables v and w are irrelevant, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u\,;$$

(ii) if all three categorical variables are mutually independent, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w\,;$$

(iii) if u and v are associated but are both independent of w, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv}\,;$$

(iv) if u and v are conditionally independent given w, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ik}^{uw} + \lambda_{jk}^{vw}\,; \quad \text{and}$$

(v) if there is no three-factor interaction, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw}\,,$$

which appears as the most complete submodel (or as the global model if the saturated model is not considered at all).

This representation naturally embeds log-linear modeling within a model choice perspective in that it calls for a selection of the most parsimonious submodel that remains compatible with the observations. This is clearly equivalent to a variable-selection problem of a special kind in the sense that *all* indicators related with the same association must remain or vanish *at once*. This specific feature means that there are much fewer submodels to consider than in a regular variable-selection problem.

Exercise 4.20. Given a contingency table with four categorical variables, determine the number of submodels to consider.

As stressed above, the representation (4.7) is not identifiable. Although the following is not strictly necessary from a Bayesian point of view (since the Bayesian approach can handle nonidentifiable settings and still estimate properly identifiable quantities), it is customary to impose identifiability constraints on the parameters as in the ANOVA model. A common convention is to set to zero the parameters corresponding to the first category of each variable, which is equivalent to removing the indicator (or *dummy variable*) of the first category for each variable (or group of variables). For instance, for a 2×2 contingency table with two variables u and v, both having two categories, say 1 and 2, the constraint could be

$$\lambda_1^u = \lambda_1^v = \lambda_{11}^{uv} = \lambda_{12}^{uv} = \lambda_{21}^{uv} = 0\,.$$

For notational convenience, we assume below that β is the vector of the parameters once the identifiability constraint has been applied and that X is the indicator matrix with the corresponding columns removed.

4.5.2 Inference Under a Flat Prior

Even when using a noninformative flat prior on β, $\pi(\beta) \propto 1$, the posterior distribution

$$\begin{aligned}\pi(\beta|\mathbf{y}, X) &\propto \prod_{i=1}^{n} \left\{\exp(\mathbf{x}^{i\mathsf{T}}\beta)\right\}^{y_i} \exp\{-\exp(\mathbf{x}^{i\mathsf{T}}\beta)\} \\ &= \exp\left\{\sum_{i=1}^{n} y_i\, \mathbf{x}^{i\mathsf{T}}\beta - \sum_{i=1}^{n} \exp(\mathbf{x}^{i\mathsf{T}}\beta)\right\} \\ &= \exp\left\{\left(\sum_{i=1}^{n} y_i\, \mathbf{x}^{i}\right)^{\mathsf{T}} \beta - \sum_{i=1}^{n} \exp(\mathbf{x}^{i\mathsf{T}}\beta)\right\}\end{aligned}$$

is nonstandard and must be approximated by an MCMC algorithm. While the shape of this density differs from the posterior densities in the probit and logit cases, we can once more implement Algorithm 4.2 based on the normal Fisher approximation of the likelihood (whose parameters are again derived using the R `glm()` function as in

```
> mod=summary(glm(y~-1+X,family=poisson()))
```

which provides $\hat{\beta}$ in mod$coeff[,1] and $\hat{\Sigma}$ in mod$cov.unscaled).

Table 4.1. Dataset airquality: Bayes estimates of the parameter β using a random walk MH algorithm with scale factor $\tau^2 = 0.5$.

Effect	Post. mean	Post. var.
λ	2.8041	0.0612
λ_2^u	–1.0684	0.2176
λ_2^v	–5.8652	1.7141
λ_2^w	–1.4401	0.2735
λ_3^w	–2.7178	0.7915
λ_4^w	–1.1031	0.2295
λ_5^w	–0.0036	0.1127
λ_{22}^{uv}	3.3559	0.4490
λ_{22}^{uw}	-1.6242	1.2869
λ_{23}^{uw}	–0.3456	0.8432
λ_{24}^{uw}	–0.2473	0.6658
λ_{25}^{uw}	–1.3335	0.7115
λ_{22}^{vw}	4.5493	2.1997
λ_{23}^{vw}	6.8479	2.5881
λ_{24}^{vw}	4.6557	1.7201
λ_{25}^{vw}	3.9558	1.7128

⊞ For **airquality**, we first consider the most general nonsaturated model, as described in Section 4.5.1. Taking into account the identifiability constraints, there are therefore

$$1+(2-1)+(2-1)+(5-1)+(2-1)\times(2-1)+(2-1)\times(5-1)+(2-1)\times(5-1)\,,$$

i.e., 16, free parameters in the model (to be compared with the 20 counts in the contingency table). Given the dimension of the simulated parameter, it is impossible to provide a complete picture of the convergence properties of the algorithm, and we represented in Figure 4.8 the traces and histograms for the marginal posterior distributions of the parameters β_i based on $10,000$ iterations using a scale factor equal to $\tau^2 = 0.5$. (This value was obtained by trial and error, producing a smooth trace for all parameters. Larger values of τ required a larger number of iterations since the acceptance rate was lower, as the reader can check using the R programs available on the book's Website.) Note that some of the traces in Figure 4.8 show some periodic patterns that indicate that more iterations could be necessary. However, the corresponding histograms remain quite stable over iterations. Both the approximated posterior means and the posterior variances for the 16 parameters as deduced from the MCMC run are given in Table 4.1. Note that a few histograms are centered at 0, signaling a potential lack of significance for the corresponding β_i's.

4.5.3 Model Choice and Significance of the Parameters

If we try to compare different levels of association (or interaction), or if we simply want to test the significance of some parameters β_i, the flat prior is once again inappropriate. The G-prior alternative proposed for the probit and logit models is still available, though, and we can thus replace the posterior distribution of the previous section with

$$\pi(\beta|\mathbf{y},X) \propto |X^{\mathsf{T}}X|^{1/2}\Gamma((2k-1)/4)\left(\beta^{\mathsf{T}}(X^{\mathsf{T}}X)\beta\right)^{-(2k-1)/4}\pi^{-k/2} \times \exp\left\{\left(\sum_{i=1}^{n} y_i\,\mathbf{x}^i\right)^{\mathsf{T}}\beta - \sum_{i=1}^{n}\exp(\mathbf{x}^{i\mathsf{T}}\beta)\right\} \tag{4.8}$$

as an alternative posterior.

Exercise 4.21. Find sufficient conditions on $(\mathbf{y}, X)$ for this posterior distribution to be proper.

⊞ For **airquality** and the same model as in the previous analysis, namely the maximum nonsaturated model with 16 parameters, Algorithm 4.2 can be used with (4.8) as target and $\tau^2 = 0.5$ as the scale in the random walk. The result of this simulation over $10,000$ iterations is presented in Figure 4.9. The traces of the components of β show the same slow mixing as in Figure 4.8, with similar occurrences of large deviances from the mean value that may indicate the weak identifiability of some of these parameters. Note also that the histograms of the posterior marginal distributions are rather close to those associated with the flat prior, as shown in Figure 4.8. The MCMC approximations to the

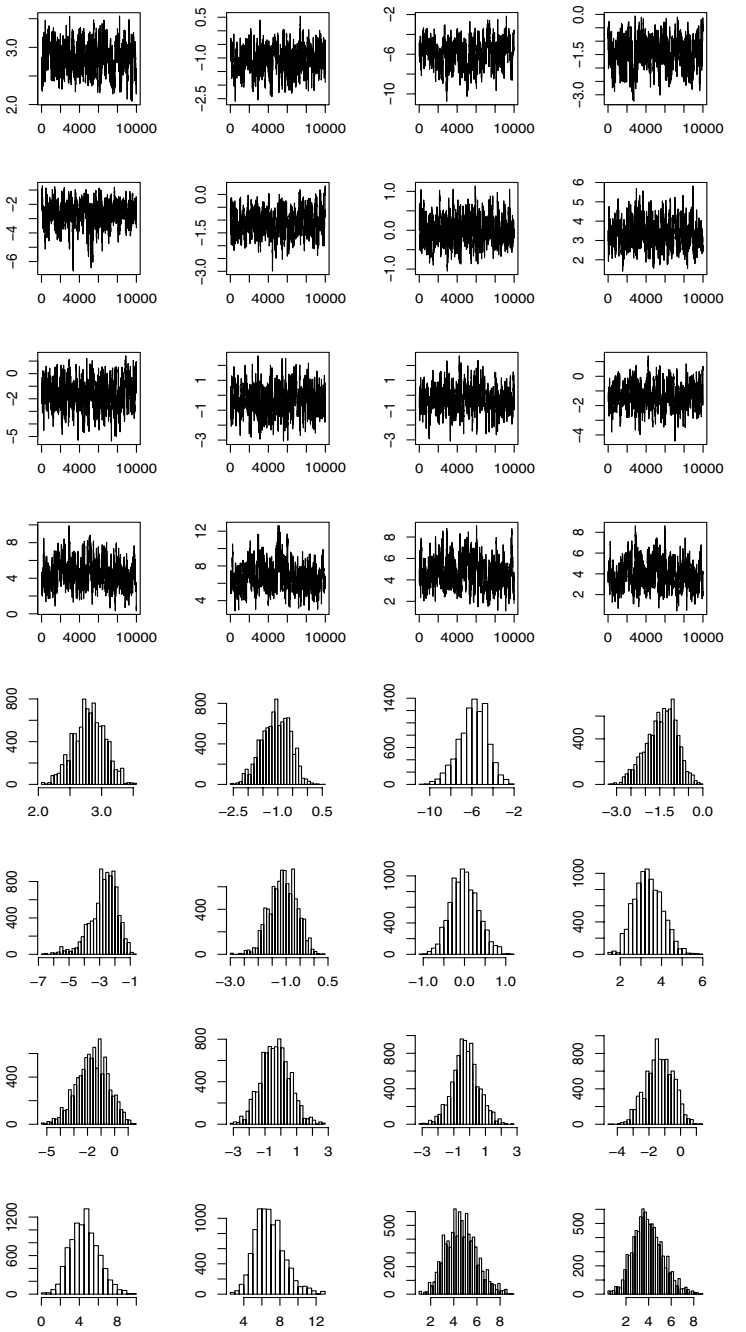


Fig. 4.8. Dataset airquality: Traces (*top*) and histograms (*bottom*) of the simulations from the posterior distributions of the components of β using a flat prior and a random walk Metropolis–Hastings algorithm with scale factor $\tau^2 = 0.5$ (same order row-wise as in Table 4.1).

posterior means and the posterior variances are given in Table 4.2 for all 16 parameters, based on the last 9000 iterations. While the first parameters are quite close to those provided by Table 4.1, the estimates of the interaction coefficients vary much more and are associated with much larger variances. This indicates that much less information is available within the contingency table about interactions, as can be expected.

Table 4.2. Dataset airquality: Metropolis–Hastings approximations of the posterior means under the G-prior.

Effect	Post. mean	Post. var.
λ	2.7202	0.0603
λ_2^u	–1.1237	0.1981
λ_2^v	–4.5393	0.9336
λ_2^w	–1.4245	0.3164
λ_3^w	–2.5970	0.5596
λ_4^w	–1.1373	0.2301
λ_5^w	0.0359	0.1166
λ_{22}^{uv}	2.8902	0.3221
λ_{22}^{uw}	–0.9385	0.8804
λ_{23}^{uw}	0.1942	0.6055
λ_{24}^{uw}	0.0589	0.5345
λ_{25}^{uw}	–1.0534	0.5220
λ_{22}^{vw}	3.2351	1.3664
λ_{23}^{vw}	5.3978	1.3506
λ_{24}^{vw}	3.5831	1.0452
λ_{25}^{vw}	2.8051	1.0061

If we now consider the very reason why this alternative to the flat prior was introduced, we are facing the same difficulty as in the probit case for the computation of the marginal density of $\mathbf{y}$. And, once again, the same solution applies: using an importance sampling experiment to approximate the integral works when the importance function is a multivariate normal (or t) distribution with mean (approximately) $\mathbb{E}[\beta|\mathbf{y}, X]$ and covariance matrix (approximately) $2 \times \mathbb{V}(\beta|\mathbf{y}, X)$ using the Metropolis–Hastings approximations reported in Table 4.2. We can therefore approximate Bayes factors for testing all possible structures of the log-linear model.

⊞ For **airquality**, we illustrate this ability by testing the presence of two-by-two interactions between the three variables. We thus compare the largest non-saturated model with each submodel where one interaction is removed. An ANOVA-like output is

```
Effect log10(BF)

u:v     6.0983 (****)
u:w    -0.5732
v:w     6.0802 (****)
```

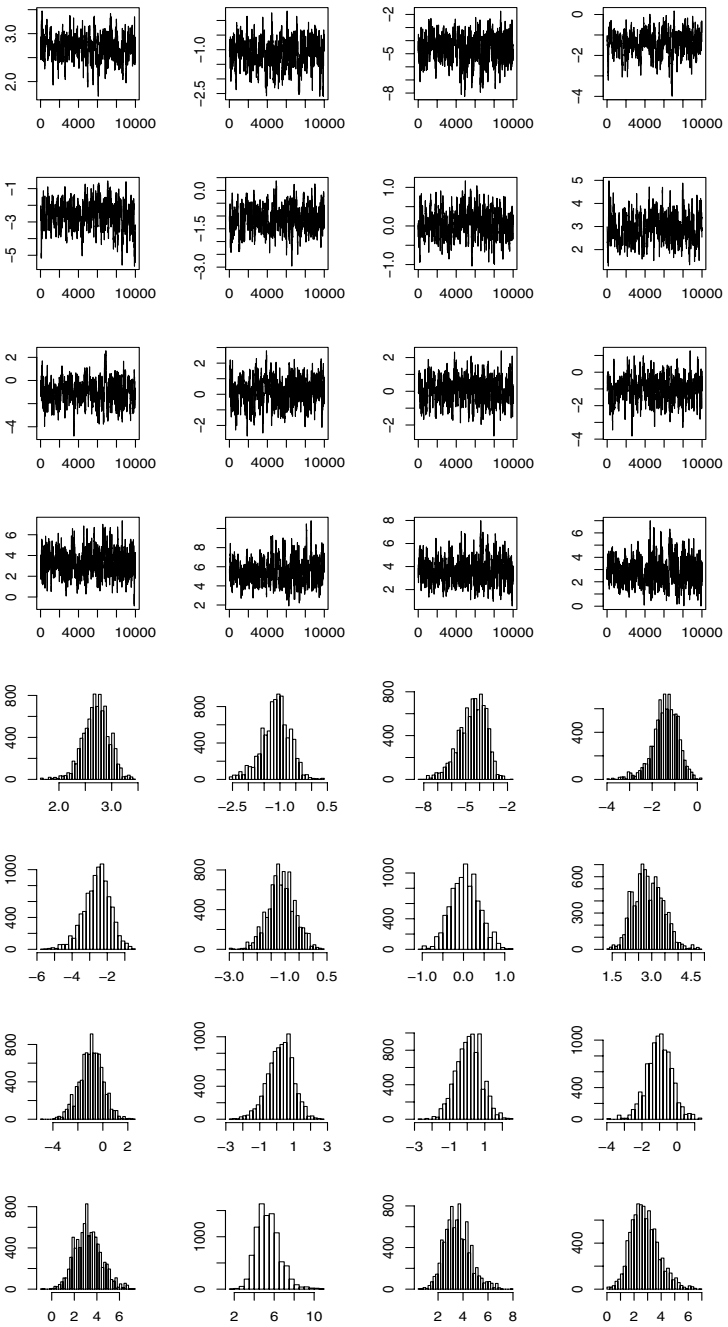


Fig. 4.9. Dataset airquality: Same legend as Figure 4.8 for the posterior distribution (4.8) as target.

```
evidence against H0: (****) decisive, (***) strong,
(**) substantial, (*) poor
```

which means that the interaction between u and w (that is, ozone and month) is too small to be significant given all the other effects. (Note that it would be excessive to derive from this lack of significance a conclusion of independence between u and w because this interaction is conditional on all other interactions in the complete nonsaturated model.)

5

Capture–Recapture Experiments

He still couldn't be sure that
he hadn't landed in a trap.
—**Ian Rankin**, ***Resurrection Men***.—

Roadmap

This chapter deals with a special case of survey models. Surveys are used in many settings to evaluate some features of a given population, including its main characteristic; that is, the *size* of the population. In the case of capture–recapture surveys, individuals are observed either once or several times and the repeated observations can be used to draw inference on the population size and its dynamic characteristics. Along with the original model, we will also introduce extensions that are a first entry into hidden Markov chain models, detailed further in Chapter 6. In particular, we cover the generic *Arnason–Schwarz* model that is customarily used for open populations.

On the methodological side, we provide an entry into the accept–reject method, which is the central simulation technique behind most standard random generators and relates to the Metropolis–Hastings methodology in many ways.

5.1 Inference in a Finite Population

In this chapter, we consider the problem of estimating the unknown size, N, of a population, based on a *survey*; that is, on a partial observation of this population. To be able to evaluate a population size without going through the enumeration of all its members is obviously very appealing, both timewise and moneywise, especially when sampling those members has a perturbing effect on them. (In the most extreme cases, sampling an individual may lead to its destruction, as for instance in forestry when estimating the volume of trees or in meat production when estimating the content of fat in meat.)

A primary type of survey (which we do not study in this chapter) is based on a knowledge of the structure of the population. For instance, in a political survey about voting intentions, we build a sample of 1000 individuals, say, such that the main sociological groups (farmers, civil servants, senior citizens, etc.) are represented in proportion in the sample. In that situation, there is no statistical inference, so to speak, except about the variability of the responses, which are in the simplest cases binomial variables.

Obviously, such surveys require primary knowledge of the population, which can be obtained either by a (costly) census, like those that states run every five or ten years, or by a preliminary exploratory survey that aims at uncovering these hidden structures. This secondary type of survey is the purpose of this chapter, under the name of *capture–recapture* (or *capture–mark–recapture*) experiments, where a few individuals sampled at random from the population of interest bring some information about the characteristics of this population and in particular about its size.

The capture–recapture models were first used in biology and ecology to estimate the size of animal populations, such as herds of caribous (for culling), families of whales (for the International Whaling Commission to determine fishing quotas), cod populations, and the number of species in a particular area. While our illustrative dataset will be related to a biological problem, we stress that those capture–recapture models apply in a much wider range of domains, such as, for instance,

- sociology and demography, where the estimation of the size of populations at risk is always delicate (e.g., homeless people, prostitutes, illegal migrants, drug addicts, etc.);
- official statistics for reducing the cost of a census[1] or improving its efficiency on delicate or rare subcategories (as in the U.S. census undercount procedure and the new French census);
- finance (e.g., in credit scoring, defaulting companies, etc.) and marketing (consumer habits, telemarketing, etc.);

[1] Even though a census is formally a deterministic process since it aims at the complete enumeration of a given population, it inevitably involves many random components at the selection, collection, and processing levels (Särndal et al., 2003).

– fraud detection (e.g., phone, credit card, etc.) and document authentication (historical documents, forgery, etc.); and
– software debugging, to determine an evaluation of the number of bugs in a computer program.

In these different examples, the size N of the whole population is unknown but samples (with fixed or random sizes) can easily be extracted from the population. For instance, in a computer program, the total number N of bugs is unknown but one can record the number n_1 of bugs detected in a given perusal. Similarly, the total number N of homeless people in a city like New York at a given time is not known but it is possible to count the number n_1 of homeless persons in a given shelter on a precise night, to record their id, and to cross this sample with a sample of n_2 persons collected the night after in order to detect how many persons n_{11} were present in the shelter on both nights.

In conclusion, we hope that the introduction above was motivating enough to convince the reader that population sampling models are deeply relevant in statistical practice. Besides, those models also provide an interesting application of Bayesian modeling and in particular they allow for the inclusion of often available prior information.

5.2 Sampling Models

5.2.1 The Binomial Capture Model

We start with the simplest model of all, namely the independent observation or *capture*[2] of n^+ individuals from a population of size N. For instance, a trap is positioned on a rabbit track for five hours and n^+ rabbits are found in the trap. While the population size $N \in \mathbb{N}^*$ is the parameter of interest, there exists a nuisance parameter, namely the probability $p \in [0,1]$ with which each individual is captured. (This model assumes that catching the ith individual is independent of catching the jth individual.) For this model,

$$n^+ \sim \mathscr{B}(N,p)$$

and the corresponding likelihood is

$$\ell(N,p|n^+) = \binom{N}{n^+} p^{n^+}(1-p)^{N-n^+}\mathbb{I}_{N\geq n^+}\,.$$

Obviously, with a single observation n^+, we cannot say much on (N,p), but the posterior distribution is still well-defined. For instance, if we use the vague prior

[2]We use the original terminology of *capture* and *individuals*, even though the sampling mechanism may be far from genuine capture, as in whale sightseeing or software bug detection.

$$\pi(N,p) \propto N^{-1}\mathbb{I}_{\mathbb{N}^*}(N)\mathbb{I}_{[0,1]}(p)\,,$$

the posterior distribution of N is

$$\begin{aligned}\pi(N|n^+) &\propto \frac{N!}{(N-n^+)!}\,N^{-1}\mathbb{I}_{N\geq n^+}\mathbb{I}_{\mathbb{N}^*}(N)\int_0^1 p^{n^+}(1-p)^{N-n^+}\,\mathrm{d}p\\ &\propto \frac{(N-1)!}{(N-n^+)!}\,\frac{(N-n^+)!}{(N+1)!}\,\mathbb{I}_{N\geq n^+\vee 1}\\ &= \frac{1}{N(N+1)}\,\mathbb{I}_{N\geq n^+\vee 1}\,, \end{aligned} \tag{5.1}$$

where $n^+ \vee 1 = \max(n^+, 1)$. Note that this posterior distribution is defined even when $n^+ = 0$. If we use the (more informative) uniform prior

$$\pi(N,p) \propto \mathbb{I}_{\{1,\ldots,S\}}(N)\mathbb{I}_{[0,1]}(p)\,,$$

the posterior distribution of N is

$$\pi(N|n^+) \propto \frac{1}{N+1}\mathbb{I}_{\{n^+\vee 1,\ldots,S\}}(N)\,.$$

Exercise 5.1. Show that the posterior distribution $\pi(N|n^+)$ given by (5.1), while associated with an improper prior, is defined for all values of n^+. Show that the normalization factor of (5.1) is $n^+ \vee 1$, and deduce that the posterior median is equal to $2(n^+ \vee 1)$. Discuss the relevance of this estimator and show that it corresponds to a Bayes estimate of p equal to $1/2$.

Exercise 5.2. Under the same prior, derive the marginal posterior density of N in the case where $n_1^+ \sim \mathscr{B}(N,p)$ and

$$n_2^+, \ldots, n_k^+ \overset{\text{iid}}{\sim} \mathscr{B}(n_1^+, p)$$

are observed (the later are in fact recaptures). Apply to the sample

$$(n_1^+, n_2^+, \ldots, n_{11}^+) = (32, 20, 8, 5, 1, 2, 0, 2, 1, 1, 0)\,,$$

which describes a series of tag recoveries over 11 years.

⊞ The dataset we consider throughout this chapter is called **eurodip** and is related to a population of birds called *European dippers (Cinclus cinclus)*. These birds are closely dependent on streams, feeding on underwater invertebrates, and their nests are always close to water. The capture–recapture data on the European dipper contained in **eurodip** covers 7 years (1981–1987 inclusive) of observations in a zone of 200 km^2 in eastern France. The data consists of markings and recaptures of breeding adults each year during the breeding period from early March to early June. Birds were at least one year old

when initially banded. In **eurodip**, each row of seven digits corresponds to a capture–recapture story for a given dipper, 0 indicating an absence of capture that year and, in the case of a capture, 1,2, or 3 representing the zone where the dipper is captured. For instance, the lines

```
1 0 0 0 0 0 0
1 3 0 0 0 0 0
0 2 2 2 1 2 2
```

indicate that the first dipper was only captured the first year in zone 1 and that the second dipper was captured in years 1981 and 1982 and moved from zone 1 to zone 3 between those years. The third dipper was captured every year but in 1981 and moved between zones 1 and 2 during the remaining year.

For illustrative purposes, consider the case of year 1981 (which is the first column in **eurodip**), where $n^+ = 22$ dippers were captured. By using the binomial capture model and the vague prior $\pi(N, p) \propto N^{-1}$, the number of dippers N can be estimated by the posterior median 44. (Note that the mean of (5.1) does not exist, no matter what n^+ is.) If we use the ecological information that there cannot be more than 400 dippers in this region, we can take the prior $\pi(N, p) \propto \mathbb{I}_{\{1,\ldots,400\}}(N)\mathbb{I}_{[0,1]}(p)$ and estimate the number of dippers N by its posterior expectation, 130.52.

5.2.2 The Two-Stage Capture–Recapture Model

A logical extension to the capture model above is the *capture–mark–recapture* model, which considers two capture periods plus a marking stage, as follows:

(i) n_1 individuals from a population of size N are "captured", that is, sampled without replacement.
(ii) Those individuals are "marked", that is, identified by a numbered tag (for birds and fishes), a collar (for mammals), or another device (like the Social Security number for homeless people or a picture for whales), and they are then released into the population.
(iii) A second and similar sampling (once again without replacement) is conducted, with n_2 individuals captured.
(iv) m_2 individuals out of the n_2's bear the identification mark and are thus characterized as having been captured in both experiments.

If we assume a *closed population* (that is, a fixed population size N throughout the capture experiment), a constant capture probability p for all individuals, and complete independence between individuals and between captures, we end up with a product of binomial models,

$$n_1 \sim \mathscr{B}(N, p)\,, \quad m_2|n_1 \sim \mathscr{B}(n_1, p)\,,$$

and

$$n_2 - m_2|n_1, m_2 \sim \mathscr{B}(N - n_1, p)\,.$$

If

$$n^c = n_1 + n_2 \quad \text{and} \quad n^+ = n_1 + (n_2 - m_2)$$

denote the total number of captures over both periods and the total number of captured individuals, respectively, the corresponding likelihood $\ell(N, p|n_1, n_2, m_2)$ is

$$\begin{aligned}
&\binom{N-n_1}{n_2-m_2} p^{n_2-m_2}(1-p)^{N-n_1-n_2+m_2} \mathbb{I}_{\{0,\ldots,N-n_1\}}(n_2-m_2) \\
&\qquad \times \binom{n_1}{m_2} p^{m_2}(1-p)^{n_1-m_2} \binom{N}{n_1} p^{n_1}(1-p)^{N-n_1} \mathbb{I}_{\{0,\ldots,N\}}(n_1) \\
&\qquad \propto \frac{N!}{(N-n_1-n_2+m_2)!} p^{n_1+n_2}(1-p)^{2N-n_1-n_2} \mathbb{I}_{N \geq n^+} \\
&\qquad \propto \binom{N}{n^+} p^{n^c}(1-p)^{2N-n^c} \mathbb{I}_{N \geq n^+} \,,
\end{aligned}$$

which shows that (n^c, n^+) is a sufficient statistic. If we choose the prior $\pi(N, p) = \pi(N)\pi(p)$ such that $\pi(p)$ is a $\mathscr{U}([0,1])$ density, the conditional posterior distribution on p is such that

$$\pi(p|N, n_1, n_2, m_2) = \pi(p|N, n^c) \propto p^{n^c}(1-p)^{2N-n^c} \,;$$

that is,

$$p|N, n^c \sim \mathscr{B}e(n^c + 1, 2N - n^c + 1).$$

Unfortunately, the marginal posterior distribution of N is more complicated. For instance, if $\pi(N) = \mathbb{I}_{\mathbb{N}^*}(N)$, it satisfies

$$\pi(N|n_1, n_2, m_2) = \pi(N|n^c, n^+) \propto \binom{N}{n^+} B(n^c + 1, 2N - n^c + 1) \mathbb{I}_{N \geq n^+ \vee 1} \,.$$

This distribution is called a *beta-Pascal* distribution, but it is not very tractable. The same difficulty occurs if $\pi(N) = N^{-1}\mathbb{I}_{\mathbb{N}^*}(N)$.

Ⓒ The intractability in the posterior distribution $\pi(N|n_1, n_2, m_2)$ is due to the infinite summation resulting from the unbounded support of N. A feasible approximation is to replace the missing normalizing factor by a finite sum with a large enough bound on N, the bound being determined by a lack of perceivable impact on the sum. But the approximation errors due to the computations of terms such as $\binom{N}{n^+}$ or $B(n^c + 1, 2N - n^c + 1)$ can become a serious problem when n^+ is large.

If we have information about an upper bound S on N and use the corresponding uniform prior,

$$\pi(N) \propto \mathbb{I}_{\{1,\ldots,S\}}(N) \,,$$

the posterior distribution of N is thus proportional to

$$\pi(N|n^+) \propto \binom{N}{n^+} \frac{\Gamma(2N - n^c + 1)}{\Gamma(2N + 2)} \mathbb{I}_{\{n^+ \vee 1,\ldots,S\}}(N) \,,$$

and, in this case, it is possible to calculate the posterior expectation of N with no approximation error.

⊞ For the first two years of the **eurodip** experiment, which correspond to the first two columns and the first 70 rows of the dataset, $n_1 = 22$, $n_2 = 59$, and $m_2 = 11$. Therefore, within the frame of the two-stage capture–recapture model[3] and the uniform prior $\mathscr{U}(\{1,\ldots,400\}) \times \mathscr{U}([0,1])$ on (N,p), the posterior expectation of N is equal to 165.26.

A simpler model used in capture–recapture settings is the hypergeometric model, also called the *Darroch model*. This model can be seen as a conditional version of the two-stage model when conditioning on both sample sizes n_1 and n_2 since

$$m_2|n_1,n_2 \sim \mathscr{H}(N,n_2,n_1/N)\,. \tag{5.2}$$

If we choose the uniform prior $\mathscr{U}(\{1,\ldots,400\})$ on N, the posterior distribution of N is thus

$$\pi(N|m_2) \propto \binom{N-n_1}{n_2-m_2} \Big/ \binom{N}{n_2} \mathbb{I}_{\{n^+ \vee 1,\ldots,400\}}(N)\,,$$

and posterior expectations can be computed numerically by simple summations.

Exercise 5.3. Show that the conditional distribution of m_2 conditional on both sample sizes n_1 and n_2 is given by (5.2) and does not depend on p. Deduce the expectation $\mathbb{E}[m_2|n_1,n_2,N]$.

⊞ For the first two years of the **eurodip** dataset and $S = 400$, the posterior distribution of N for the Darroch model is given by

$$\pi(N|m_2) \propto (n-n_1)!(N-n_2)!/\{(n-n_1-n_2+m_2)!N!\}\,\mathbb{I}_{\{71,\ldots,400\}}(N)\,,$$

the normalization factor being the inverse of

$$\sum_{k=71}^{400} (k-n_1)!(k-n_2)!/\{(k-n_1-n_2+m_2)!k!\}\,.$$

We thus have a closed-form posterior distribution and the posterior expectation of N is equal to 137.60. Table 5.1 shows the evolution of this posterior expectation for different values of m_2. The number of recaptures is thus highly influential on the estimate of N. In parallel, Table 5.2 shows the evolution of the posterior expectation for different values of S. When S is large enough, say larger than $S = 250$, the estimate of N is quite stable, as expected.

[3] This analysis is based on the assumption that all birds captured in the second year were already present in the population during the first year.

Table 5.1. Dataset eurodip: Rounded posterior expectation of the dipper population size, N, under a uniform prior $\mathscr{U}(\{1,\ldots,400\})$.

m_2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\mathbb{E}[N\|m_2]$	355	349	340	329	316	299	277	252	224	197	172	152	135	122	110	101

Table 5.2. Dataset eurodip: Rounded posterior expectation of the dipper population size, N, under a uniform prior $\mathscr{U}(\{1,\ldots,S\})$, for $m_2 = 11$.

S	100	150	200	250	300	350	400	450	500
$\mathbb{E}[N\|m_2]$	95	125	141	148	151	151	152	152	152

Exercise 5.4. In order to determine the number N of buses in a town, a capture–recapture strategy goes as follows. We observe $n_1 = 20$ buses during the first day and keep track of their identifying numbers. Then we repeat the experiment the following day by recording the number of buses that have already been spotted on the previous day, say $m_2 = 5$, out of the $n_2 = 30$ buses observed the second day. For the Darroch model, give the posterior expectation of N under the prior $\pi(N) = 1/N$.

Exercise 5.5. Show that the maximum likelihood estimator of N for the Darroch model is $\hat{N} = n_1/(m_2/n_2)$, and deduce that it is not defined when $m_2 = 0$.

Exercise 5.6. Give the likelihood of the extension of Darroch's model when the capture–recapture experiments are repeated K times with capture sizes and recapture observations n_k $(1 \le k \le K)$ and m_k $(2 \le k \le K)$, respectively. (*Hint*: Exhibit first the two-dimensional sufficient statistic associated with this model.)

Leaving the Darroch model and getting back to the two-stage capture model with probability p of capture, the posterior distribution of (N,p) associated with the noninformative prior $\pi(N,p) = 1/N$ is proportional to

$$\frac{(N-1)!}{(N-n^+)!}\, p^{n^c}(1-p)^{2N-n^c}\,.$$

Thus, if $n^+ > 0$, both conditional posterior distributions are standard distributions since

$$\begin{aligned} p|n^c, N &\sim \mathscr{B}e(n^c+1, 2N-n^c+1) \\ N-n^+|n^+, p &\sim \mathscr{N}eg(n^+, 1-(1-p)^2)\,. \end{aligned}$$

Indeed, as a function of N,

$$\frac{(N-1)!}{(N-n^+)!}\,(1-p)^{2N-n^c} \propto \binom{N-1}{N-n^+} \left\{(1-p)^2\right\}^{N-n^+} \left\{1-(1-p)^2\right\}^{n^+}\,.$$

Therefore, while the marginal posterior in N is difficult to manage, the joint distribution of (N, p) can be approximated by a Gibbs sampler, as follows:

Algorithm 5.1. Two-stage Capture–Recapture Gibbs Sampler
Initialization: Generate $p^{(0)} \sim \mathscr{U}([0,1])$.
Iteration t $(t \geq 1)$:
1. Generate $N^{(t)} - n^+ \sim \mathscr{N}eg(n^+, 1-(1-p^{(t-1)})^2)$.
2. Generate $p^{(t)} \sim \mathscr{B}e(n^c+1, 2N^{(t)} - n^c + 1)$.

Exercise 5.7. Give both conditional posterior distributions in the case $n^+ = 0$.

Exercise 5.8. Show that, when the prior on N is a $\mathscr{P}(\lambda)$ distribution, the conditional posterior on $N - n_+$ is $\mathscr{P}(\lambda(1-p)^2)$.

5.2.3 The T-Stage Capture–Recapture Model

A further extension to the two-stage capture–recapture model is to consider instead a series of T consecutive captures. In that case, if we denote by n_t the number of individuals captured at period t $(1 \leq t \leq T)$ and by m_t the number of recaptured individuals (with the convention that $m_1 = 0$), under the same assumptions as in the two-stage model, then $n_1 \sim \mathscr{B}(N, p)$ and, conditionally on the $j-1$ previous captures and recaptures $(2 \leq j \leq T)$,

$$m_j \sim \mathscr{B}\left(\sum_{t=1}^{j-1}(n_t - m_t), p\right) \quad \text{and} \quad n_j - m_j \sim \mathscr{B}\left(N - \sum_{t=1}^{j-1}(n_t - m_t), p\right).$$

The likelihood $\ell(N, p | n_1, n_2, m_2 \ldots, n_T, m_T)$ is thus

$$\begin{aligned}
&\binom{N}{n_1} p^{n_1}(1-p)^{N-n_1} \prod_{j=2}^{T} \left[\binom{N - \sum_{t=1}^{j-1}(n_t - m_t)}{n_j - m_j} p^{n_j - m_j + m_j} \right. \\
&\qquad \left. \times (1-p)^{N - \sum_{t=1}^{j}(n_t - m_t)} \binom{\sum_{t=1}^{j-1}(n_t - m_t)}{m_j} (1-p)^{\sum_{t=1}^{j-1}(n_t - m_t) - m_j} \right] \\
&\qquad \propto \frac{N!}{(N-n^+)!} p^{n^c} (1-p)^{TN - n^c} \mathbb{I}_{N \geq n^+}
\end{aligned}$$

if we denote the sufficient statistics as

$$n^+ = \sum_{t=1}^{T}(n_t - m_t) \quad \text{and} \quad n^c = \sum_{t=1}^{T} n_t \,,$$

the total numbers of captured individuals and captures over the T periods, respectively.

For a noninformative prior such as $\pi(N,p) = 1/N$, the joint posterior satisfies

$$\pi(N,p|n^+,n^c) \propto \frac{(N-1)!}{(N-n^+)!}\, p^{n^c}(1-p)^{TN-n^c}\mathbb{I}_{N\geq n^+\vee 1}\,.$$

Therefore, the conditional posterior distribution of p is

$$p|N,n^+,n^c \sim \mathscr{B}e(n^c+1, TN-n^c+1)$$

and the marginal posterior distribution of N

$$\pi(N|n^+,n^c) \propto \frac{(N-1)!}{(N-n^+)!}\,\frac{(TN-n^c)!}{(TN+1)!}\,\mathbb{I}_{N\geq n^+\vee 1}\,,$$

is computable. Note that the normalization coefficient can also be approximated by summation with an arbitrary precision unless N and n^+ are very large.

For the uniform prior $\mathscr{U}(\{1,\ldots,S\})$ on N and $\mathscr{U}([0,1])$ on p, the posterior distribution of N is thus proportional to

$$\pi(N|n^+) \propto \binom{N}{n^+}\frac{(TN-n^c)!}{(TN+1)!}\,\mathbb{I}_{\{n^+\vee 1,\ldots,S\}}(N).$$

⊞ For the whole set of observations in **eurodip**, we have $T=7$, $n^+=294$, and $n^c=519$. Under the uniform prior with $S=400$, the posterior expectation of N is equal to 372.89. While this value seems dangerously close to the upper bound of 400 on N and thus leads us to suspect a strong influence of the upper bound S, the computation of the posterior expectation for $S=2500$ leads to 373.99, which shows the limited impact of this hyperparameter.

Using even a slightly more advanced sampling model may lead to genuine computational difficulties. For instance, consider a heterogeneous capture–recapture model where the individuals are captured at time $1 \leq t \leq T$ with probability p_t and where both the size N of the population and the probabilities p_t are unknown. The corresponding likelihood is

$$\ell(N,p_1,\ldots,p_T|n_1,n_2,m_2\ldots,n_T,m_T) \propto \frac{N!}{(N-n^+)!}\prod_{t=1}^{T} p_t^{n_t}(1-p_t)^{N-n_t}.$$

If the associated prior on $(N,p_1,\ldots,p_T)$ is such that

$$N \sim \mathscr{P}(\lambda)$$

and $(1 \leq t \leq T)$,

$$\alpha_t = \log\left(\frac{p_t}{1-p_t}\right) \sim \mathcal{N}(\mu_t, \sigma^2),$$

where both σ^2 and the μ_t's are known,[4] the posterior distribution satisfies

$$\pi(\alpha_1, \ldots, \alpha_T, N|, n_1, \ldots, n_T) \propto \frac{N!}{(N-n^+)!}\,\frac{\lambda^N}{N!}\,\prod_{t=1}^{T}(1+e^{\alpha_t})^{-N} \times \prod_{t=1}^{T}\exp\left\{\alpha_t n_t - \frac{1}{2\sigma^2}(\alpha_t-\mu_t)^2\right\}. \tag{5.3}$$

It is thus much less manageable from a computational point of view, especially when there are many capture episodes. A corresponding Gibbs sampler could simulate easily from the conditional posterior distribution on N since

$$N - n^+|\alpha, n^+ \sim \mathcal{P}oi\left(\lambda\prod_{t=1}^{T}(1+e^{\alpha_t})\right),$$

but the conditionals on the α_t's $(1 \le t \le T)$ are less conventional,

$$\alpha_t|N, \mathbf{n} \sim \pi_t(\alpha_t|N, \mathbf{n}) \propto (1+e^{\alpha_t})^{-N} e^{\alpha_t n_t - (\alpha_t-\mu_t)^2/2\sigma^2},$$

and they require either an accept–reject algorithm (Section 5.4) or a Metropolis–Hastings algorithm in order to be simulated.

Exercise 5.9. An extension of the T-stage capture–recapture model is to consider that the capture of an individual modifies its probability of being captured from p to q for future recaptures. Give the likelihood $\ell(N, p, q|n_1, n_2, m_2 \ldots, n_T, m_T)$.

Exercise 5.10. Another extension of the 2-stage capture–recapture model is to allow for mark losses.[5] If we introduce q as the probability of losing the mark, r as the probability of recovering a lost mark and k as the number of recovered lost marks, give the associated likelihood $\ell(N, p, q, r|n_1, n_2, m_2, k)$.

For the prior

$$\pi(N, P) \propto \frac{\lambda^N}{N!}\mathbb{I}_{\mathbb{N}}(N)\mathbb{I}_{[0,1]}(p),$$

[4] This assumption can be justified on the basis that each capture probability is only observed once on the tth round (and so cannot reasonably be associated with a noninformative prior).

[5] Tags can be lost by marked animals, but the animals themselves could also be lost to recapture either by changing habitat or dying. Our current model assumes that the population is *closed*; that is, that there is no immigration, emigration, birth, or death within the population during the length of the study. This other kind of extension will be dealt with in Sections 5.3 and 5.5.

the conditional posteriors are then

$$p|N, n^c \sim \mathscr{B}(n^c + 1, TN - n^c + 1) \quad \text{and} \quad N - n^+|p, n^+ \sim \mathscr{P}(\lambda(1-p)^T)$$

and a Gibbs sampler similar to the one developed in Algorithm 5.1 can easily be implemented.

⊞ For **eurodip**, we used this Gibbs sampler and obtained the results illustrated by Figure 5.1. When the chain is initialized at the (unlikely) value $N^{(0)} = \lambda = 200$ (which is the prior expectation of N), the stabilization of the chain is quite clear: It only takes a few iterations to converge toward the proper region that supports the posterior distribution. We can thus visually confirm the convergence of the algorithm and approximate the Bayes estimator of N by 327.01 and the Bayes estimator of p by 0.23 (which are the Monte Carlo averages). The precision of these estimates can be assessed as in a regular Monte Carlo experiment, but the variance estimate is biased because of the correlation between the simulations.

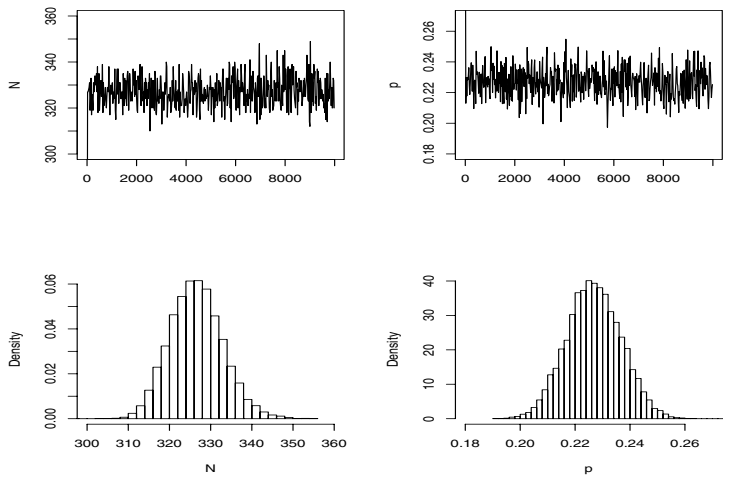


Fig. 5.1. Dataset eurodip: Representation of the Gibbs sampling output for the parameters p (*first column*) and N (*second column*).

Exercise 5.11. Reproduce the analysis of eurodip when switching the prior from $\pi(N, p) \propto \lambda^N/N!$ to $\pi(N, p) \propto N^{-1}$.

5.3 Open Populations

We consider now the case of an *open population* model, where the population size does not remain fixed over the experiment but, on the contrary, there is a probability q for each individual to leave the population at each time (or, more accurately, between each capture episode). Given that the associated likelihood involves unobserved indicators (namely, indicators of survival; see Exercise 5.14), we study here a simpler model where only the individuals captured during the first capture experiment are marked and subsequent recaptures are registered. For three successive capture experiments, we thus have

$$n_1 \sim \mathscr{B}(N,p)\,, \quad r_1|n_1 \sim \mathscr{B}(n_1,q)\,, \quad r_2|n_1,r_1 \sim \mathscr{B}(n_1-r_1,q)\,,$$

and

$$c_2|n_1,r_1 \sim \mathscr{B}(n_1-r_1,p), \quad c_3|n_1,r_1,r_2 \sim \mathscr{B}(n_1-r_1-r_2,p)\,,$$

where only n_1, c_2, and c_3 are observed. The numbers of individuals removed at stages 1 and 2, r_1 and r_2, are not available and must therefore be simulated, as well as the parameters N, p, and q.[6] The likelihood $\ell(N,p,q,r_1,r_2|n_1,c_2,c_3)$ is given by

$$\begin{aligned}&\binom{N}{n_1}p^{n_1}(1-p)^{N-n_1}\binom{n_1}{r_1}q^{r_1}(1-q)^{n_1-r_1}\binom{n_1-r_1}{c_2}p^{c_2}(1-p)^{n_1-r_1-c_2}\\ &\quad\times\binom{n_1-r_1}{r_2}q^{r_2}(1-q)^{n_1-r_1-r_2}\binom{n_1-r_1-r_2}{c_3}p^{c_3}(1-p)^{n_1-r_1-r_2-c_3}\end{aligned}$$

and, if we use the prior $\pi(N,p,q) = N^{-1}\mathbb{I}_{[0,1]}(p)\mathbb{I}_{[0,1]}(q)$, the associated conditionals are

$$\begin{aligned}\pi(p|N,q,\mathcal{D}^*) &\propto p^{n_+}(1-p)^{u_+}\,,\\ \pi(q|N,p,\mathcal{D}^*) &\propto q^{c_1+c_2}(1-q)^{2n_1-2r_1-r_2}\,,\\ \pi(N|p,q,\mathcal{D}^*) &\propto \frac{(N-1)!}{(N-n_1)!}(1-p)^N\mathbb{I}_{N\geq n_1}\,,\\ \pi(r_1|p,q,n_1,c_2,c_3,r_2) &\propto \frac{(n_1-r_1)!\,q^{r_1}(1-q)^{-2r_1}(1-p)^{-2r_1}}{r_1!(n_1-r_1-r_2-c_3)!(n_1-c_2-r_1)!}\,,\\ \pi(r_2|p,q,n_1,c_2,c_3,r_1) &\propto \frac{q^{r_2}[(1-p)(1-q)]^{-r_2}}{r_2!(n_1-r_1-r_2-c_3)!}\,,\end{aligned}$$

where $\mathcal{D}^* = (n_1,c_2,c_3,r_1,r_2)$ and

[6] From a theoretical point of view, r_1 and r_2 are *missing variables* rather than true parameters. This obviously does not change anything either for simulation purposes or for Bayesian inference.

$$u_1 = N - n_1,\ u_2 = n_1 - r_1 - c_2,\ u_3 = n_1 - r_1 - r_2 - c_3\,,$$
$$n_+ = n_1 + c_2 + c_3,\ u_+ = u_1 + u_2 + u_3$$

(u stands for *unobserved*). Therefore, the full conditionals are

$$\begin{aligned} p|N,q,\mathcal{D}^* &\sim \mathcal{B}e(n_+ + 1, u_+ + 1)\,,\\ q|N,p,\mathcal{D}^* &\sim \mathcal{B}e(r_1 + r_2 + 1, 2n_1 - 2r_1 - r_2 + 1)\,,\\ N - n_1|p,q,\mathcal{D}^* &\sim \mathcal{N}eg(n_1,p)\,,\\ r_2|p,q,n_1,c_2,c_3,r_1 &\sim \mathcal{B}\left(n_1 - r_1 - c_3, \frac{q}{1+(1-q)(1-p)}\right)\,, \end{aligned}$$

which are very easily simulated, while r_1 has a less conventional distribution. However, this difficulty is minor since, in our case, n_1 is not extremely large. It is thus possible to compute the probability that r_1 is equal to one of the values in $\{0, 1, \ldots, \min(n_1 - r_2 - c_3, n_1 - c_2)\}$.

Ⓒ We stress that R is quite helpful in simulating from unusual distributions and in particular from those with finite support. For instance, the conditional distribution of r_1 above can be simulated using the following representation of $\mathbb{P}(r_1 = k|p,q,n_1,c_2,c_3,r_2)$ $(0 \le k \le \overline{r} = \min(n_1 - r_2 - c_3, n_1 - c_2))$,

$$\binom{n_1 - c_2}{k} \left\{\frac{q}{(1-q)^2(1-p)^2}\right\}^k \binom{n_1 - k}{r_2 + c_3}\,, \tag{5.4}$$

up to a normalization constant, since the binomial coefficients and the power in k can be computed for all values of k at once, due to the matrix capabilities of R, through the command `lchoose`. The quantity

```
pr=lchoose(n=n1 - c2,k=0:r̄) + (0:r̄)*log(q1)
          + lchoose(n=n1-(0:r̄),k=r2 + c3)
```

is the whole vector of the log-probabilities, with $q_1 = q/(1-q)^2(1-p)^2$.

↯ In most computations, it is safer to use logarithmic transforms to reduce the risk of running into overflow or underflow error messages. For instance, in the example above, the probability vector can be recovered by

```
pr=exp(pr-max(pr))/sum(exp(pr-max(pr)))
```

while a direct computation of `exp(pr)` may well produce an `Inf` value that invalidates the remaining computations.[7]

Ⓒ Once the probabilities are transformed as in the previous R code, a call to the R command

[7] This recommendation also applies to the computation of likelihoods that tend to take absolute values that exceed the range of the computer representation of real numbers, while only the relative values are relevant for Bayesian computations. Using a transform such as `exp(loglike-max(loglike))` thus helps in reducing the risk of overflows.

```
> sample(1:rbar,n)
```

is sufficient to provide n simulations of r_1. Even a large value such as $n_1 = 1612$ used below does not lead to computing difficulties since we can run $10,000$ iterations of the corresponding Gibbs sampler in less than 2 minutes on a laptop.

⊞ For **eurodip**, we have $n_1 = 22$, $c_2 = 11$, and $c_3 = 6$. We obtain the Gibbs output summarized in Figure 5.2 (in that case, the starting value for N was 200, while r_1 and r_2 were started at 10 and 5, respectively). The sequences for all components are rather stable and their mixing behavior (i.e., the speed of exploration of the support of the target) is satisfactory, even though we can still detect a trend in the first three rows. Since r_1 and r_2 are integers with only a few possible values, the last two rows show apparently higher jumps than the three other parameters. The MCMC approximations to the posterior expectations of N and p are equal to 57 and 0.40, respectively.

Given the large difference between n_1 and c_2 and the proximity between c_2 and c_3, high values of q are rejected, and the difference can be attributed with high likelihood to a poor capture rate. One should take into account the fact that there are only three observations for a model that involves three true parameters plus two missing variables. Figure 5.3 gives another insight into the posterior distribution by representing the joint distribution of the sample of (r_1, r_2)'s using for representation purposes the R function jitter(), which moves each point by a tiny random amount. There is a clear positive correlation between r_1 and r_2, despite the fact that r_2 is simulated on an $(n_1 - c_3 - r_1)$ scale. The mode of the posterior is $(r_1, r_2) = (0, 0)$, which means that it is likely that no dipper died or left the observation area over the three-year period.

Exercise 5.12. Show that the conditional distribution of r_1 is indeed proportional to the product (5.4).

Exercise 5.13. Show that r_2 can be integrated out in the joint distribution above and leads to the following distribution on r_1:

$$\pi(r_1|p,q,n_1,c_2,c_3) \propto \frac{(n_1-r_1)!(n_1-r_1-c_3)!}{r_1!(n_1-r_1-c_2)!} \times \left(\frac{q}{(1-p)(1-q)[q+(1-p)(1-q)]}\right)^{r_1}.$$

Compare the computational cost of a Gibbs sampler based on this approach with a Gibbs sampler using the full conditionals.

Exercise 5.14. Show that the likelihood associated with an open population can be written as

$$\ell(N,p|\mathcal{D}^*) = \sum_{(\epsilon_{it},\delta_{it})_{it}} \prod_{t=1}^{T}\prod_{i=1}^{N} q_{\epsilon_{i(t-1)}}^{\epsilon_{it}}(1-q_{\epsilon_{i(t-1)}})^{1-\epsilon_{it}} \times p^{(1-\epsilon_{it})\delta_{it}}(1-p)^{(1-\epsilon_{it})(1-\delta_{it})},$$

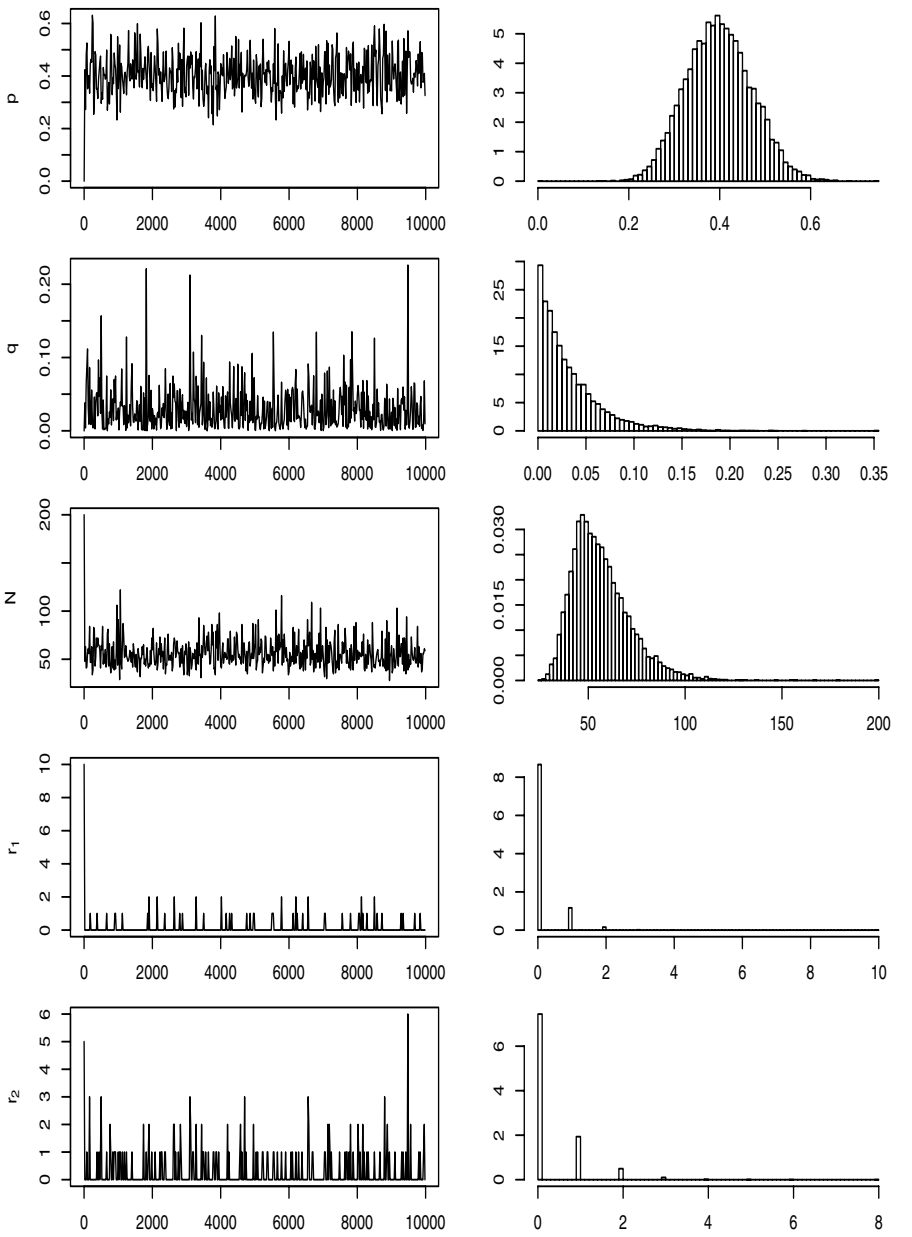


Fig. 5.2. Dataset eurodip: Representation of the Gibbs sampling output for the five parameters of the open population model, based on 10,000 iterations, with raw plots *(first column)* and histograms *(second column)*.

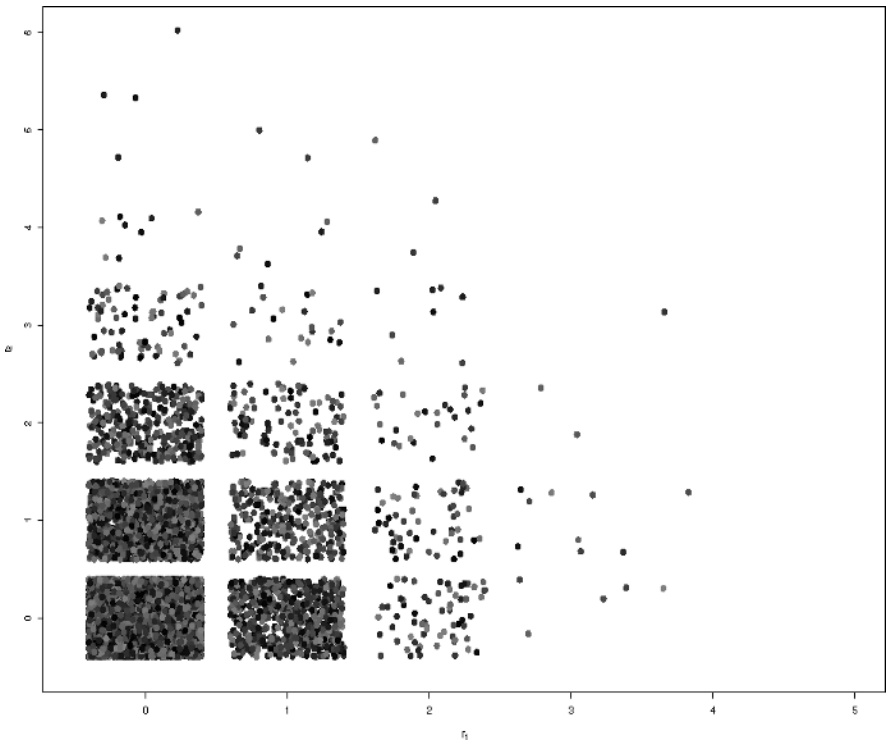


Fig. 5.3. Dataset eurodip: Representation of the Gibbs sampling output of the (r_1, r_2)'s by a jitterplot: To translate the density of the possible values of (r_1, r_2) on the $\mathbb{N}^2$ grid, each simulation has been randomly moved using the R jitter procedure and colored at random using grey levels to help distinguish the various simulations.

where $q_0 = q$, $q_1 = 1$, and δ_{it} and ϵ_{it} are the capture and exit indicators, respectively. Derive the order of complexity of this likelihood; that is, the number of elementary operations necessary to compute it.[8]

5.4 Accept–Reject Algorithms

In Chapter 2, we mentioned standard random number generators used for the most common distributions and presented importance sampling (Algorithm 2.2) as a possible alternative when such generators are not available. While

[8] We will see in Chapter 7 a derivation of this likelihood that enjoys an $O(T)$ complexity.

MCMC algorithms always offer a solution when facing nonstandard distributions, there often exists a possibility that is used in most of the standard random generators. It also relates to the independent Metropolis–Hastings algorithm of Section 4.2.2.

Given a density g that is defined on an arbitrary space (of whatever dimension), a fundamental identity is that simulating X distributed from $g(x)$ is completely equivalent to simulating (X, U) uniformly distributed on the set

$$\mathscr{S} = \{(x, u) : 0 < u < g(x)\}$$

(this is called the *Fundamental Theorem of Simulation* in Robert and Casella, 2004, Chapter 3). The reason for this equivalence is simply that

$$\int_0^\infty \mathbb{I}_{0<u<g(x)}\, \mathrm{d}u = g(x)\,.$$

Exercise 5.15. Show that, for $M > 0$, if g is replaced with Mg in $\mathscr{S}$ and if (X, U) is uniformly distributed on $\mathscr{S}$, the marginal distribution of X is still g. Deduce that the density g only needs to be known up to a normalizing constant.

Exercise 5.16. For the function $g(x) = (1+\sin^2(x))(2+\cos^4(4x))\exp[-x^4\{1+\sin^6(x)\}]$ on $[0, 2\pi]$, examine the feasibility of running a uniform sampler on the associated set $\mathscr{S}$.

Since $\mathscr{S}$ usually has complex features, direct simulation from the uniform distribution on $\mathscr{S}$ is most often impossible (Exercise 5.16). The idea behind the accept–reject method is to find a simpler set $\mathscr{G}$ that contains $\mathscr{S}$, $\mathscr{S} \subset \mathscr{G}$, and then to simulate uniformly on this set $\mathscr{G}$ until the value belongs to $\mathscr{S}$. In practice, this means that one needs to find an upper bound on g; that is, another density f and a constant M such that

$$g(x) \le Mf(x) \tag{5.5}$$

on the support of the density g. (Note that $M > 1$ necessarily.) Implementing the following algorithm then leads to a simulation from g.

Algorithm 5.2. Accept–Reject Sampler
1. Generate $X \sim f$, $U \sim \mathcal{U}_{[0,1]}$.
2. Accept $Y = x$ if $u \le g(x)/(Mf(x))$.
3. Return to 1 otherwise.

Note that this method provides a random generator for densities g that are known up to a multiplicative factor, which is a feature that occurs particularly often in Bayesian calculations since the posterior distribution is usually specified up to a normalizing constant.

Exercise 5.17. Show that the probability of acceptance in Step 2 of Algorithm 5.2 is $1/M$ and that the number of trials until a variable is accepted has a geometric distribution with parameter $1/M$. Conclude that the expected number of trials per simulation is M.

Exercise 5.18. For the conditional distribution of α_t derived from (5.3), construct an accept–reject algorithm based on a normal bounding density f and study its performances for $N = 532$, $n_t = 118$, $\mu_t = -0.5$, and $\sigma^2 = 3$.

Exercise 5.19. When uniform simulation on $\mathscr{S}$ is impossible, construct a Gibbs sampler based on the conditional distributions of u and x. (*Hint*: Show that both conditionals are uniform distributions.) This special case of the Gibbs sampler is called the *slice sampler* (see Robert and Casella, 2004, Chapter 8). Apply to the distribution of Exercise 5.16.

⊞ For the open population model, we found the full conditional distribution of r_1 to be non-standard, as shown by (5.4). Rather than using an exhaustive enumeration of all probabilities $\mathbb{P}(m_1 = k) = g(k)$ and then sampling from this distribution, we can instead try to use a proposal based on a binomial upper bound. If f corresponds to the binomial distribution $\mathscr{B}(n_1, q_1)$ ($q_1 = q/(1-q)^2(1-p)^2$), the ratio $g(k)/f(k)$ is proportional to

$$\frac{\binom{n_1-c_2}{k}(1-q_1)^k\binom{n_1-k}{r_2+c_3}}{\binom{n_1}{k}} = \frac{(n_1-c_2)!}{(r_2+c_3)!n_1!}\,\frac{((n_1-k)!)^2(1-q_1)^k}{(n_1-c_2-k)!(n_1-r_2-c_3-k)!}$$

and the second ratio is decreasing in k. The ratio is therefore bounded by

$$\frac{(n_1-c_2)!}{(r_2+c_3)!}\,\frac{n_1!}{(n_1-c_2)!(n_1-r_2-c_3)!} = \binom{n_1}{r_2+c_3}.$$

Note that this is *not* the constant M because we use unnormalized densities (the bound M may therefore also depend on q_1). Therefore we cannot derive the average acceptance rate from this ratio and we have to use a Monte Carlo experiment to check whether or not the method is really efficient.

If we use the values in **eurodip**—that is, $n_1 = 22$, $c_2 = 11$ and $c_3 = 6$, with $r_2 = 1$ and $q_1 = 0.1$—the average of the acceptance ratios $g(k)/Mf(k)$ is equal to 0.12. This is a relatively small value since it corresponds to a rejection rate of about $9/10$. The simulation process could thus be a little slow. A histogram of accepted values is shown in Figure 5.4.

Exercise 5.20. Reproduce the analysis above for the marginal distribution of r_1 computed in Exercise 5.13.

Obviously, this method is not hassle-free. For complex densities g, it may prove impossible to find a density f such that $g(x) \leq Mf(x)$ and M is small

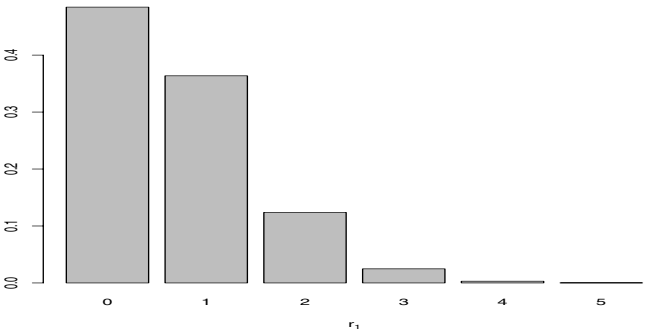


Fig. 5.4. Dataset eurodip: Sample from the distribution (5.4) obtained by accept–reject and based on the simulation of 10,000 values from a $\mathscr{B}(n_1, q_1)$ distribution for $n_1 = 22$, $c_2 = 11$, $c_3 = 6$, $r_2 = 1$, and $q_1 = 0.1$.

enough. However, there exists a large class of univariate distributions for which a generic choice of f is possible (see Robert and Casella, 2004, Chapter 2).

5.5 The Arnason–Schwarz Capture–Recapture Model

We consider in this last section a more advanced capture–recapture model based on the realistic assumption that, in most capture–recapture experiments, we can tag individuals one by one; that is, we can distinguish each individual at the time of its first capture and thus follow its capture history. For instance, when tagging mammals and birds, differentiated tags can be used, so that there is only *one* individual with tag 23131932.[9]

The *Arnason–Schwarz model* thus considers a capture–recapture experiment as a collection of individual histories. For each individual that has been captured at least once during the experiment, individual characteristics of interest are registered at each capture. For instance, this may include location, weight, sexual status, pregnancy occurrence, social status, and so on. The probabilistic modeling includes this categorical decomposition by adding what we will call *movement* probabilities to the survival probabilities already used in the Darroch open population model of Section 3.4.3. From a theoretical point of view, this is a first example of a (partially) hidden Markov model,

[9] In a capture–recapture experiment used in Dupuis (1995), a population of lizards was observed in the south of France (Lozère). When it was found that plastic tags caused necrosis on those lizards, the biologists in charge of the experiment decided to cut a phalange of one of the fingers of the captured lizards to identify them later. While the number of possibilities, 2^{10}, is limited, it is still much larger than the number of captured lizards in this study. Whether or not the lizards appreciated this ability to classify them is not known.

a structure studied in detail in Chapter 7. In addition, the model includes the possibility that individuals vanish from the population between two capture experiments. (This is thus another example of an open population model.)

As in eurodip, the interest that drives the capture–recapture experiment may be to study the movements of individuals within a zone $\mathfrak{K}$ divided into $k = 3$ strata denoted by $1, 2, 3$. (This structure is generic: Zones are not necessarily geographic and can correspond to anything from social status, to HIV stage, to university degree.) For instance, four consecutive rows of possible eurodip (individual) capture–recapture histories look as follows:

45	0 3 0 0 0 0 0
46	0 2 2 2 2 1 1
47	0 2 0 0 0 0 0
48	2 1 2 1 0 0 0

where 0 denotes a failure to capture. This means that, for dipper number 46, the first location was not observed but this dipper was captured for all the other experiments. For dipper numbers 45 and 47, there was no capture after the second time and thus one or both of them could be dead (or outside the range of the capture area) at the time of the last capture experiment. We also stress that the Arnason–Schwarz model often assumes that individuals that were not there on the first capture experiments can be identified.[10] We thus have *cohorts* of individuals that entered the study in the first year, the second year, and so on.

5.5.1 Modeling

A description of the model involves two types of variables for each individual i $(i = 1, \ldots, n)$ in the population: first, a variable that describes the location of this individual,

$$\mathbf{z}_i = (z_{(i,t)}, t = 1, .., \tau)\,,$$

where τ is the number of capture periods; and, second, a binary variable that describes the capture history of this individual,

$$\mathbf{x}_i = (x_{(i,t)}, t = 1, .., \tau)\,.$$

We assume that $z_{(i,t)} = r$ when the animal i is alive in stratum r at time t and that $z_{(i,t)} = \dagger$ denotes the case when the animal i is dead at time t. The variable $\mathbf{z}_i$ is sometimes called the *migration* process of individual i by analogy with the special case where one is considering animals moving between geographical zones, like some northern birds in spring and fall. Note that $\mathbf{x}_i$ is entirely observed, while $\mathbf{z}_i$ is not. For instance, we may have

$$\mathbf{x}_i = 1\,1\,0\,1\,1\,1\,0\,0\,0$$

[10] This is the case, for instance, with newborns or new mothers in animal capture experiments.

and

$$\mathbf{z}_i = 1\ 2\ \cdot\ 3\ 1\ 1\ \cdot\ \cdots,$$

for which a possible $\mathbf{z}_i$ is

$$\mathbf{z}_i = 1\ 2\ 1\ 3\ 1\ 1\ 2\ \dagger\ \dagger\,,$$

meaning that the animal died between the seventh and the eighth capture events. In particular, the Arnason–Schwarz model assumes that dead animals are never observed (although this type of assumption can easily be modified when processing the model, in what are called *tag-recovery experiments*). Therefore $z_{(i,t)} = \dagger$ always corresponds to $x_{(i,t)} = 0$. Moreover, we assume that the $(\mathbf{x}_i, \mathbf{z}_i)$'s $(i = 1, \ldots, n)$ are independent and that each random vector $\mathbf{z}_i$ is a Markov chain taking values in $\mathfrak{K} \cup \{\dagger\}$ with uniform initial probability on $\mathfrak{K}$. The parameters of the Arnason–Schwarz model are thus the capture probabilities

$$p_t(r) = \mathbb{P}\left(x_{(i,t)} = 1 | z_{(i,t)} = r\right)$$

on the one hand and the transition probabilities

$$q_t(r,s) = \mathbb{P}\left(z_{(i,t+1)} = s | z_{(i,t)} = r\right) \quad r \in \mathfrak{K}, s \in \mathfrak{K} \cup \{\dagger\}, \quad q_t(\dagger,\dagger) = 1$$

on the other hand. We denote by $\varphi_t(r) = 1 - q_t(r,\dagger)$ the *survival* probabilities and by $\psi_t(r,s)$ the interstrata *movement* probabilities, defined as

$$q_t(r,s) = \varphi_t(r) \times \psi_t(r,s) \quad r \in \mathfrak{K}, s \in \mathfrak{K}\,.$$

The likelihood $\ell(p_1, \ldots, p_\tau, q_1, \ldots, q_\tau | (\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n))$ is then given by

$$\prod_{i=1}^{n} \left[\prod_{t=1}^{\tau} p_t(z_{(i,t)})^{x_{(i,t)}} (1 - p_t(z_{(i,t)}))^{1-x_{(i,t)}} \times \prod_{t=1}^{\tau-1} q_t(z_{(i,t)}, z_{(i,t+1)}) \right], \quad (5.6)$$

up to a constant. The prior modeling corresponding to those parameters will depend on the information that is available about the population covered by the capture–recapture experiment. For illustration's sake, consider the use of conjugate priors

$$p_t(r) \sim \mathscr{B}e(a_t(r), b_t(r))\,, \qquad \varphi_t(r) \sim \mathscr{B}e(\alpha_t(r), \beta_t(r))\,,$$

where the hyperparameters $a_t(r)$ and so on depend on both time t and location r, and

$$\psi_t(r) \sim \mathscr{D}ir(\gamma_t(r))\,,$$

where $\psi_t(r) = (\psi_t(r,s); s \in \mathfrak{K})$ with

$$\sum_{s \in \mathfrak{K}} \psi_t(r,s) = 1\,,$$

and $\gamma_t(r) = (\gamma_t(r,s); s \in \mathfrak{K})$. The determination of these (numerous) hyperparameters is also case-dependent and varies from a noninformative modeling,

where all hyperparameters are taken to be equal to 1 or 1/2, to a very informative setting where exact values of these hyperparameters can be chosen from the prior information. The following example is an illustration of the latter.

Table 5.3. Prior information about the capture and survival parameters of the Arnason–Schwarz model, represented by prior expectation and prior confidence interval, for a capture–recapture experiment on the migrations of lizards. (*Source*: Dupuis, 1995.)

Episode	2	3	4	5	6
p_t Mean	0.3	0.4	0.5	0.2	0.2
95% cred. int.	$[0.1, 0.5]$	$[0.2, 0.6]$	$[0.3, 0.7]$	$[0.05, 0.4]$	$[0.05, 0.4]$

Site	A		B,C	
Episode	$t = 1, 3, 5$	$t = 2, 4$	$t = 1, 3, 5$	$t = 2, 4$
$\varphi_t(r)$ Mean	0.7	0.65	0.7	0.7
95% cred. int.	$[0.4, 0.95]$	$[0.35, 0.9]$	$[0.4, 0.95]$	$[0.4, 0.95]$

Example 5.1. For the capture–recapture experiment described in footnote 9 on the migrations of lizards between three adjacent zones, there are six capture episodes. The prior information provided by the biologists on the capture and survival probabilities, p_t (which are assumed to be zone independent) and $\varphi_t(r)$, is given by Table 5.3. While this may seem very artificial, this construction of the prior distribution actually happened that way because the biologists in charge were able to quantify their beliefs and intuitions in terms of prior expectation and prior confidence interval. (The differences in the prior values on p_t are due to differences in capture efforts, while the differences between episodes 1, 3 and 5, and episodes 2 and 4 are due to the fact that the odd indices correspond to spring and the even indices to fall and mortality is higher over the winter.) Moreover, this prior information can be perfectly translated in a collection of beta priors (see Exercise 5.21) whose hyperparameters are given in Table 5.4. ◀

Table 5.4. Hyperparameters of the beta priors corresponding to the information contained in Table 5.3. (*Source*: Dupuis, 1995.)

Episode	2	3	4	5	6
Dist.	$\mathscr{B}e(6, 14)$	$\mathscr{B}e(8, 12)$	$\mathscr{B}e(12, 12)$	$\mathscr{B}e(3.5, 14)$	$\mathscr{B}e(3.5, 14)$

Site	A		B	
Episode	$t = 1, 3, 5$	$t = 2, 4$	$t = 1, 3, 5$	$t = 2, 4$
Dist.	$\mathscr{B}e(6.0, 2.5)$	$\mathscr{B}e(6.5, 3.5)$	$\mathscr{B}e(6.0, 2.5)$	$\mathscr{B}e(6.0, 2.5)$

Exercise 5.21. Show that, given a mean and a 95% confidence interval in $[0,1]$, there exists at most one beta distribution $\mathscr{B}(a,b)$ with such a mean and confidence interval.

5.5.2 Gibbs Sampler

A Gibbs sampler for the Arnason–Schwarz model needs to account for the missing components in the vectors $\mathbf{z}_i$ in order to simulate the parameters from the full conditional distribution

$$\pi(\theta|\mathbf{x},\mathbf{z}) \propto \ell(\theta|\mathbf{x},\mathbf{z}) \times \pi(\theta)\,,$$

where $\mathbf{x}$ and $\mathbf{z}$ denote the collections of the vectors of capture indicators and locations, respectively. This is thus a particular case of *data augmentation*, where the missing data $\mathbf{z}$ are simulated at each step t in order to reconstitute a complete sample $(\mathbf{x},\mathbf{z}^{(t)})$ for which conjugacy applies. If we can simulate the full conditional distributions both of the parameters and of the missing components, the Gibbs sampler is as follows:

Algorithm 5.3. Arnason–Schwarz Gibbs Sampler
Iteration l $(l \geq 1)$:

1. **Parameter simulation**
   Simulate $\theta^{(l)} \sim \pi(\theta|\mathbf{z}^{(l-1)},\mathbf{x})$ as $(t=1,\ldots,\tau)$,

$$p_t^{(l)}(r)|\mathbf{x},\mathbf{z}^{(l-1)} \sim \mathscr{B}e\left(a_t(r)+u_t(r), b_t(r)+v_t^{(l)}(r)\right)\,,$$

$$\varphi_t^{(l)}(r)|\mathbf{x},\mathbf{z}^{(l-1)} \sim \mathscr{B}e\left(\alpha_t(r)+\sum_{j\in\mathfrak{K}} w_t^{(l)}(r,j), \beta_t(r)+w_t^{(l)}(r,\dagger)\right)\,,$$

$$\psi_t^{(l)}(r)|\mathbf{x},\mathbf{z}^{(l-1)} \sim \mathscr{D}ir\left(\gamma_t(r,s)+w_t^{(l)}(r,s); s\in\mathfrak{K}\right)\,,$$

   where

$$w_t^{(l)}(r,s) = \sum_{i=1}^{n} \mathbb{I}_{(z_{(i,t)}^{(l-1)}=r, z_{(i,t+1)}^{(l-1)}=s)}\,,$$

$$u_t^{(l)}(r) = \sum_{i=1}^{n} \mathbb{I}_{(x_{(i,t)}=1, z_{(i,t)}^{(l-1)}=r)}\,,$$

$$v_t^{(l)}(r) = \sum_{i=1}^{n} \mathbb{I}_{(x_{(i,t)}=0, z_{(i,t)}^{(l-1)}=r)}\,.$$

2. **Missing location simulation**
Generate the unobserved $z_{(i,t)}^{(l)}$'s from the full conditional distributions

$$\mathbb{P}(z_{(i,1)}^{(l)} = s|x_{(i,1)}, z_{(i,2)}^{(l-1)}, \theta^{(l)}) \propto q_1^{(l)}(s, z_{(i,2)}^{(l-1)})(1 - p_1^{(l)}(s))\,,$$
$$\mathbb{P}(z_{(i,t)}^{(l)} = s|x_{(i,t)}, z_{(i,t-1)}^{(l)}, z_{(i,t+1)}^{(l-1)}, \theta^{(l)}) \propto q_{t-1}^{(l)}(z_{(i,t-1)}^{(l)}, s) \times q_t(s, z_{(i,t+1)}^{(l-1)})(1 - p_t^{(l)}(s))\,,$$
$$\mathbb{P}(z_{(i,\tau)}^{(l)} = s|x_{(i,\tau)}, z_{(i,\tau-1)}^{(l)}, \theta^{(l)}) \propto q_{\tau-1}^{(l)}(z_{(i,\tau-1)}^{(l)}, s)(1 - p_\tau(s)^{(l)})\,.$$

Ⓒ Simulating the missing locations in the $\mathbf{z}_i$'s conditionally on the other locations is not a very complex task because of the good conditioning properties of these vectors (which stem from their Markovian nature). As already implemented in Step 2. of Algorithm 5.3, the full conditional distribution of $z_{(i,t)}$ only depends on the previous and next locations $z_{(i,t-1)}$ and $z_{(i,t+1)}$ (and obviously on the fact that it is not observed; that is, that $x_{(i,t)} = 0$). When the number of states $s \in \mathfrak{K}$ is moderate, it is straightforward to simulate from such a distribution.

Take $\mathfrak{K} = \{1, 2\}$, $n = 4$, $m = 8$ and assume that, for $\mathbf{x}$, we have the following histories:

1	1	1	·	·	1	·	·	·
2	1	·	1	·	1	·	2	1
3	2	1	·	1	2	·	·	1
4	1	·	·	1	2	1	1	2

Assume also that all (prior) hyperparameters are taken equal to 1. Then one possible instance of a simulated $\mathbf{z}$ is

1	1	1	2	1	1	2	†
1	1	1	2	1	1	1	2
2	1	2	1	2	1	1	1
1	2	1	1	2	1	1	2

and it leads to the following simulation of the parameters:

$$p_4^{(l)}(1)|\mathbf{x}, \mathbf{z}^{(l-1)} \sim \mathscr{B}e(1 + 2, 1 + 0)\,,$$
$$\varphi_7^{(l)}(2)|\mathbf{x}, \mathbf{z}^{(l-1)} \sim \mathscr{B}e(1 + 0, 1 + 1)\,,$$
$$\psi_2^{(l)}(1, 2)|\mathbf{x}, \mathbf{z}^{(l-1)} \sim \mathscr{B}e(1 + 1, 1 + 2)\,,$$

in the Gibbs sampler, where the hyperparameters are therefore derived from the (partly) simulated history above. Note that because there are only two possible states, the Dirichlet distribution simplifies into a beta distribution.

Exercise 5.22. Show that groups of consecutive unknown locations are independent of one another, conditional on the observations. Devise a way to simulate these groups by blocks rather than one at a time; that is, using the joint posterior distributions of the groups rather than the full conditional distributions of the states.

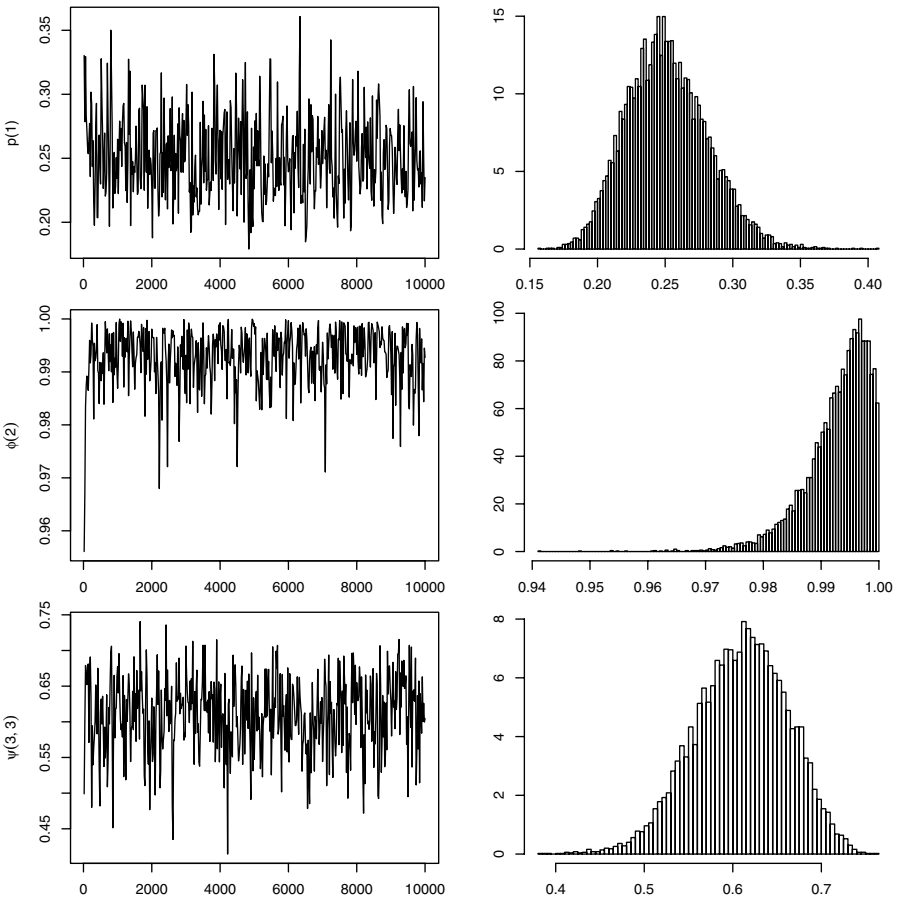


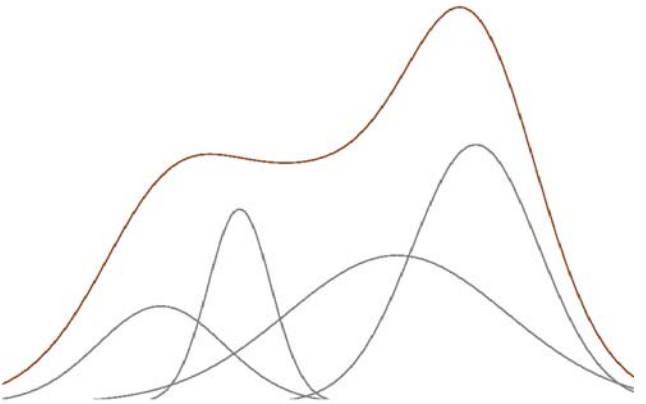
Fig. 5.5. Dataset eurodip: Representation of the Gibbs sampling output for some parameters of the Arnason–Schwarz model, based on 10,000 iterations, with raw plots (*first column*) and histograms (*second column*).

⊞ For **eurodip**, Lebreton et al. (1992) argue that the capture and survival rates should be constant over time. If we assume that the movement probabilities are also time

independent, we are left with $3 + 3 + 3 \times 2 = 12$ parameters. Figure 5.5 gives the Gibbs output for the parameters $p(1)$, $\varphi(2)$, and $\psi(3,3)$ using noninformative priors with $a(r) = b(r) = \alpha(r) = \beta(r) = \gamma(r,s) = 1$. The convergence of the Gibbs sampler to the region of interest occurs very quickly, even though we can spot an approximate periodicity in the raw plots on the left-hand side. The MCMC approximations of the estimates of $p(1)$, $\varphi(2)$, and $\psi(3,3)$, the empirical mean over the last 8000 simulations, are equal to 0.25, 0.99, and 0.61, respectively.

6

Mixture Models



I must have missed something.
—**Ian Rankin**, ***The Hanging Garden***.—

Roadmap

This chapter covers a class of models where a rather simple distribution is degraded by an attrition mechanism that mixes together several known or unknown distributions. This representation is naturally called a mixture of distributions. Inference about the parameters of the elements of the mixtures and the weights is called mixture estimation, while recovery of the original distribution of each observation is called classification (or, more exactly, unsupervised classification to distinguish it from the supervised classification to be discussed in Chapter 8). Both aspects almost always require advanced computational tools since even the representation of the posterior distribution may be complicated. Typically, Bayesian inference for these models was not correctly treated until the introduction of MCMC algorithms in the early 1990s.

This chapter is also the right place to introduce the concept of "variable dimension models," where the structure (dimension) of the model is determined a posteriori using the data. This opens new perspectives for Bayesian inference such as model averaging but calls for a special simulation algorithm called reversible jump MCMC.

6.1 Introduction

In some cases, the complexity of a model originates from the fact that some piece of information about an original and more standard (simpler) model is *missing*. For instance, we have encountered a missing variable model in Chapter 5 with the Arnason–Schwarz model (Section 5.5), where the fact of ignoring the characteristics of the individuals outside their capture periods makes inference much harder. Similarly, as we have seen in Chapter 4, the probit model can be reinterpreted as a missing-variable model in that we only observe the sign of a normal variable.

Formally, all models that are defined via a marginalization mechanism, that is, such that the density of the observables $\mathbf{x}$, $f(\mathbf{x}|\theta)$, is given by an integral

$$f(\mathbf{x}|\theta) = \int_{\mathscr{Z}} g(\mathbf{x}, \mathbf{z}|\theta)\, \mathrm{d}\mathbf{z}\,, \tag{6.1}$$

can be considered as setting a *missing variable* (or *missing data*) model.[1]

In this chapter, we focus on the case of the mixture model, which is *the* archetypical missing-variable model in that its simple representation (and interpretation) is mirrored by a need for complex processing. Later, in Chapter 7, we will also discuss *hidden Markov models* that add to the missing structure a temporal (i.e., dependent) dimension.

6.2 Finite Mixture Models

We now introduce a specific model that synthesizes the complexity of missing-variable models, both by its nature (in the sense that it is inherently linked with a missing variable) and by its processing, which also requires the incorporation of the missing structure.[2]

A *mixture distribution* is a convex combination

$$\sum_{i=1}^{k} p_i f_i(x)\,, \qquad p_i \geq 0\,, \qquad \sum_{i=1}^{k} p_i = 1\,,$$

of $k > 1$ other distributions f_i. In the simplest situation, the f_i's are known and inference focuses either on the unknown proportions p_i or on the allocations of the points of the sample to the components f_i. In most cases, however,

[1] This is not a definition in the mathematical sense since all densities can be represented that way. We thus stress that the model itself must be introduced that way. This point is not to be mistaken for a requirement that the variable $\mathbf{z}$ be meaningful for the data at hand. In many cases, for instance the probit model, the representation is completely formal.

[2] We will see later that the missing structure does not actually *need* to be simulated but, for more complex missing-variable structures such as the hidden Markov models (introduced in Chapter 7), this completion cannot be avoided.

the f_i's are from a parametric family like the normal or Beta distributions, with unknown parameters θ_i, leading to the mixture model

$$\sum_{i=1}^{k} p_i f(x|\theta_i)\,, \tag{6.2}$$

with parameters including both the weights p_i and the component parameters θ_i $(i = 1, \ldots, k)$. It is actually relevant to distinguish the weights from the other parameters in that they are solely associated with the missing-data structure of the model, while the others are related to the observations. This distinction is obviously irrelevant in the computation of the likelihood function or in the construction of the prior distribution, but it matters in the interpretation of the posterior output, for instance.

Exercise 6.1. Show that a mixture of Bernoulli distributions is again a Bernoulli distribution. Extend this to the case of multinomial distributions.

There are several motivations for considering mixtures of distributions as a useful extension to "standard" distributions. The most natural approach is to envisage a dataset as made of several latent (that is, missing, unobserved) strata or subpopulations. For instance, one of the early occurrences of mixture modeling can be found in Bertillon (1887),[3] where the bimodal structure of the heights of (military) conscripts in central France (Doubs) can be explained a posteriori by the aggregation of two populations of young men, one from the plains and one from the mountains. The mixture structure appears because the origin of each observation (that is, the allocation to a specific subpopulation or stratum) is lost. In the example of the military conscripts, this means that the geographical origin of each young man was not recorded.

Depending on the setting, the inferential goal may be either to reconstitute the groups by estimating the missing component z, an operation usually called classification (or *clustering*), to provide estimators for the parameters of the different groups, or even to estimate the number k of groups.

A completely different approach to the interpretation and estimation of mixtures is the *semiparametric* perspective. To summarize this approach, consider that since very few phenomena obey probability laws corresponding to the most standard distributions, mixtures such as (6.2) can be seen as a good trade-off between fair representation of the phenomenon and efficient estimation of the underlying distribution. If k is large enough, there is theoretical support for the argument that (6.2) provides a good approximation (in some

[3] The Frenchman Alphonse Bertillon is also the father of scientific police investigation. For instance, he originated the use of fingerprints in criminal investigations and suspect identification.

functional sense) to most distributions. Hence, a mixture distribution can be perceived as a type of basis approximation of unknown distributions, in a spirit similar to wavelets and splines, but with a more intuitive flavor (for a statistician at least). This chapter mostly focuses on the "parametric" case, when the partition of the sample into subsamples with different distributions f_j does make sense from the dataset point of view (even though the computational processing is the same in both cases).

Let us consider an iid sample $x_1, \ldots, x_n$ from model (6.2). The likelihood is such that

$$\ell(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j \, f(x_i|\theta_j) \,,$$

where $\mathbf{x} = (x_1, \ldots, x_n)$. This likelihood leads to k^n terms when the inner sums are expanded. While this expression is not necessary for computing the likelihood at a given value $(\boldsymbol{\theta}, \mathbf{p})$, which is feasible in $\mathrm{O}(nk)$ operations as demonstrated by the representation in Figure 6.1, the computational difficulty in using the expanded version precludes analytic solutions for either maximum likelihood or Bayes estimators.

Example 6.1. Consider the simple case of a two-component normal mixture

$$p\,\mathcal{N}(\mu_1, 1) + (1-p)\,\mathcal{N}(\mu_2, 1) \,, \tag{6.3}$$

where the weight $p \neq 0.5$ is known. In this case, the parameters are identifiable: μ_1 cannot be confused with μ_2 when p is different from 0.5. Nonetheless, the log-likelihood surface represented[4] in Figure 6.1 exhibits two modes, one being close to the true value of the parameters used to simulate the dataset and one corresponding to an inverse separation of the dataset into two groups.[5] ◀

For any prior $\pi(\boldsymbol{\theta}, \mathbf{p})$, the posterior distribution of $(\boldsymbol{\theta}, \mathbf{p})$ is available up to a multiplicative constant:

$$\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}) \propto \left[\prod_{i=1}^{n} \sum_{j=1}^{k} p_j \, f(x_i|\theta_j) \right] \pi(\boldsymbol{\theta}, \mathbf{p}) \,. \tag{6.4}$$

While $\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x})$ is indeed available up to a constant, and can thus be computed for a given value of $(\boldsymbol{\theta}, \mathbf{p})$ at a cost of order $\mathrm{O}(kn)$, the derivation of

[4] The R image representation of the log-likelihood is based on a discretization of the (μ_1, μ_2) space into pixels and the computation of (6.4) for the corresponding values of (μ_1, μ_2). In Figure 6.1, we used 150 values of both μ_1 and μ_2 equally spaced on $(-1.5, 4)$.

[5] To get a better understanding of this second mode, consider the limiting setting when $p = 0.5$. In that case, there are two equivalent modes of the likelihood, (μ_1, μ_2) and (μ_2, μ_1). As p moves away from 0.5, this second mode gets lower and lower compared with the other mode, but it still remains.

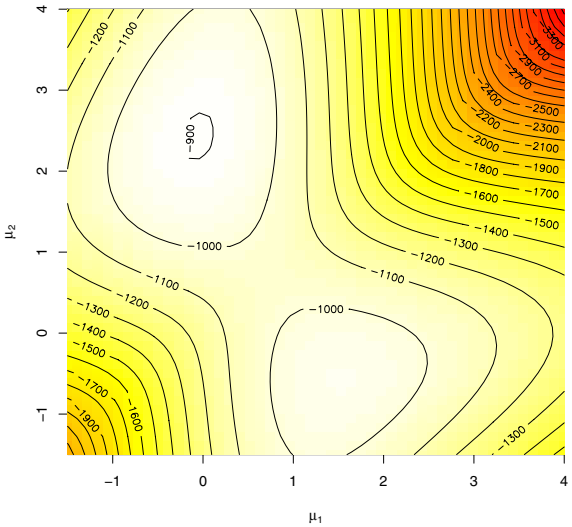


Fig. 6.1. R image representation of the log-likelihood of the mixture (6.3) for a simulated dataset of 500 observations and a true value $(\mu_1, \mu_2, p) = (0, 2.5, 0.7)$. Besides a mode in the neighborhood of the true value, the R contour function exhibits an additional mode on the likelihood surface.

the posterior characteristics, and in particular of posterior expectations of quantities of interest, is only possible in an exponential time of order $\mathrm{O}(k^n)$.

To explain this difficulty in more detail, we consider the intuitive missing-variable representation of mixture models: With each x_i is associated a missing variable z_i that indicates its component. Formally, this means that we have a hierarchical structure associated with the model:

$$z_i|\mathbf{p} \sim \mathscr{M}_k(p_1, \ldots, p_k)$$

and

$$x_i|z_i, \boldsymbol{\theta} \sim f(\cdot|\theta_{z_i})\,.$$

The completed likelihood corresponding to the missing structure is such that

$$\ell(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} p_{z_i}\, f(x_i|\theta_{z_i})$$

and

$$\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}) \propto \left[\prod_{i=1}^{n} p_{z_i}\, f(x_i|\theta_{z_i})\right] \pi\left(\boldsymbol{\theta}, \mathbf{p}\right),$$

where $\mathbf{z} = (z_1, \ldots, z_n)$. Denote by $\mathscr{Z} = \{1, \ldots, k\}^n$ the set of the k^n possible vectors $\mathbf{z}$. We can decompose $\mathscr{Z}$ into a partition of sets

$$\mathcal{Z} = \cup_{j=1}^{\mathfrak{r}} \mathcal{Z}_j$$

as follows: For a given *allocation size vector* $(n_1, \ldots, n_k)$, where $n_1 + \ldots + n_k = n$, we define the *partition sets*

$$\mathcal{Z}_j = \left\{ \mathbf{z} : \sum_{i=1}^{n} \mathbb{I}_{z_i=1} = n_1, \ldots, \sum_{i=1}^{n} \mathbb{I}_{z_i=k} = n_k \right\},$$

which consist of all allocations with the given allocation size $(n_1, \ldots, n_k)$. We set the labels $j = j(n_1, \ldots, n_k)$ by using, for instance, the lexicographical ordering on the $(n_1, \ldots, n_k)$'s. (This means that $j = 1$ corresponds to $(n_1, \ldots, n_k) = (n, 0, \ldots, 0)$, $j = 2$ to $(n_1, \ldots, n_k) = (n-1, 1, \ldots, 0)$, $j = 3$ to $(n_1, \ldots, n_k) = (n-1, 0, 1, \ldots, 0)$, and so on.

Exercise 6.2. Show that the number of nonnegative integer solutions of the decomposition of n into k parts such that $n_1 + \ldots + n_k = n$ is equal to

$$\mathfrak{r} = \binom{n+k-1}{n}.$$

Deduce that the number of partition sets is of order $\mathsf{O}(n^{k-1})$.

Using this decomposition, the posterior distribution of $(\boldsymbol{\theta}, \mathbf{p})$ can be written in closed form as

$$\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}) = \sum_{i=1}^{\mathfrak{r}} \sum_{\mathbf{z} \in \mathcal{Z}_i} \omega(\mathbf{z}) \, \pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}), \tag{6.5}$$

where $\omega(\mathbf{z})$ represents the marginal posterior probability of the allocation $\mathbf{z}$ conditional on the observations $\mathbf{x}$ (derived by integrating out the parameters $\boldsymbol{\theta}$ and $\mathbf{p}$). With this representation, a Bayes estimator of $(\boldsymbol{\theta}, \mathbf{p})$ can also be written in closed form as

$$\sum_{i=1}^{\mathfrak{r}} \sum_{\mathbf{z} \in \mathcal{Z}_i} \omega(\mathbf{z}) \, \mathbb{E}^{\pi}[\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}].$$

Example 6.2. In the special case of model (6.3), if we take two different independent normal priors on both means,

$$\mu_1 \sim \mathcal{N}(0, 4), \quad \mu_2 \sim \mathcal{N}(2, 4),$$

the posterior weight of a given allocation $\mathbf{z}$ is

$$\begin{aligned} \omega(\mathbf{z}) \propto & \sqrt{(l + 1/4)(n - l + 1/4)}\, p^l (1-p)^{n-l} \\ & \times \exp\left\{-[(l+1/4)\hat{s}_1(\mathbf{z}) + l\{\bar{x}_1(\mathbf{z})\}^2/4]/2\right\} \\ & \times \exp\left\{-[(n-l+1/4)\hat{s}_2(\mathbf{z}) + (n-l)\{\bar{x}_2(\mathbf{z}) - 2\}^2/4]/2\right\}, \end{aligned}$$

where $\ell = \sum_{i=1}^{n} \mathbb{I}_{z_i=1}$,

$$\bar{x}_1(\mathbf{z}) = \frac{1}{l}\sum_{i=1}^{n} \mathbb{I}_{z_i=1} x_i, \quad \bar{x}_2(\mathbf{z}) = \frac{1}{n-l}\sum_{i=1}^{n} \mathbb{I}_{z_i=2} x_i\,,$$

$$\hat{s}_1(\mathbf{z}) = \sum_{i=1}^{n} \mathbb{I}_{z_i=1}\left(x_i - \bar{x}_1(\mathbf{z})\right)^2, \quad \hat{s}_2(\mathbf{z}) = \sum_{i=1}^{n} \mathbb{I}_{z_i=2}\left(x_i - \bar{x}_2(\mathbf{z})\right)^2$$

(if we set $\bar{x}_1(\mathbf{z}) = 0$ when $\ell = 0$ and $\bar{x}_2(\mathbf{z}) = 0$ when $n - \ell = 0$). ◀

The decomposition (6.5) makes a lot of sense from an inferential point of view. The posterior distribution simply considers each possible partition $\mathbf{z}$ of the dataset, then allocates a posterior probability $\omega(\mathbf{z})$ to this partition, and at last constructs a posterior distribution for the parameters conditional on this allocation. Unfortunately, the computational burden resulting from this decomposition is simply too intensive because there are k^n terms in the sum.

Exercise 6.3. For a mixture of two normal distributions with all parameters unknown,

$$p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2)\,,$$

and for the prior distribution ($j = 1, 2$)

$$\mu_j|\sigma_j \sim \mathcal{N}(\xi_j, \sigma_i^2/n_j)\,, \quad \sigma_j^2 \sim \mathcal{IG}(\nu_j/2, s_j^2/2)\,, \quad p \sim \mathcal{B}e(\alpha, \beta)\,,$$

show that

$$p|\mathbf{x}, \mathbf{z} \sim \mathcal{B}e(\alpha + \ell_1, \beta + \ell_2),$$

$$\mu_j|\sigma_j, \mathbf{x}, \mathbf{z} \sim \mathcal{N}\left(\xi_1(\mathbf{z}), \frac{\sigma_j^2}{n_j + \ell_j}\right)\,,\ \sigma_j^2|\mathbf{x}, \mathbf{z} \sim \mathcal{IG}((\nu_j + \ell_j)/2, s_j(\mathbf{z})/2)\,,$$

where ℓ_j is the number of z_i equal to j, $\bar{x}_j(\mathbf{z})$ and $\hat{s}_j(\mathbf{z})$ are the empirical mean and variance for the subsample with z_i equal to j, and

$$\xi_j(\mathbf{z}) = \frac{n_j\xi_j + \ell_j\bar{x}_j(\mathbf{z})}{n_j + \ell_j}\,, \quad s_j(\mathbf{z}) = s_j^2 + \hat{s}_j^2(\mathbf{z}) + \frac{n_j\ell_j}{n_j + \ell_j}(\xi_j - \bar{x}_j(\mathbf{z}))^2\,.$$

Compute the corresponding weight $\omega(\mathbf{z})$.

There exists a solution that overcomes this computational problem. It uses an MCMC approach that takes advantage of the missing-variable structure and removes the requirement to explore the k^n possible values of $\mathbf{z}$.

Although this is beyond the scope of the book, let us point out here that there also exists in the statistical literature a technique that predates MCMC simulation algorithms but still relates to the same missing-data structure and

completion mechanism. It is called the *EM Algorithm*[6] and consists of an iterative but deterministic sequence of "E" (for *expectation*) and "M" (for *maximization*) steps that converge to a local maximum of the likelihood. At iteration t, the "E" step corresponds to the computation of the function

$$Q\{(\boldsymbol{\theta}^{(t)}, \mathbf{p}^{(t)}), (\boldsymbol{\theta}, \mathbf{p})\} = \mathbb{E}_{(\boldsymbol{\theta}^{(t)}, \mathbf{p}^{(t)})} \left[\log \ell(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z})|\mathbf{x}\right] ,$$

where the likelihood $\ell(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z})$ is the joint distribution of $\mathbf{x}$ and $\mathbf{z}$, while the expectation is computed under the conditional distribution of $\mathbf{z}$ given $\mathbf{x}$ and the value $(\boldsymbol{\theta}^{(t)}, \mathbf{p}^{(t)})$ for the parameter. The "M" step corresponds to the maximization of $Q((\boldsymbol{\theta}^{(t)}, \mathbf{p}^{(t)}), (\boldsymbol{\theta}, \mathbf{p}))$ in $(\boldsymbol{\theta}, \mathbf{p})$, with solution $(\boldsymbol{\theta}^{(t+1)}, \mathbf{p}^{(t+1)})$. As we will see in Section 6.3, the Gibbs sampler takes advantage of exactly the same conditional distribution. Further details on EM and its Monte Carlo versions (namely, when the "E" step is not analytically feasible) are given in Robert and Casella (2004, Chapter 5).

Exercise 6.4. For the normal mixture model of Exercise 6.3, compute the function $Q(\theta_0, \theta)$ and derive both steps of the EM algorithm. Apply this algorithm to a simulated dataset and test the influence of the starting point θ_0.

⊞ Although image analysis is the topic of Chapter 8, we use in this chapter a dataset derived from an image of a license plate, called **License** and represented in Figure 6.2 (top). The actual histogram of the grey levels is concentrated on 256 values because of the poor resolution of the image, but we transformed the original data as

```
> data=jitter(data,10)
> data=log((data-min(data)+.01)/(max(data)+.01-data))
```

where `jitter` is used to randomize the dataset and the second line is a logit transform. As seen from Figure 6.2 (bottom), the resulting structure of the data is acceptable as a sample from a mixture of several normal distributions (with at least two components). We must point out at this stage that mixture modeling is often used in image smoothing but not in feature recognition, which requires spatial coherence and thus more complicated models, which will be presented in Chapter 8.

6.3 MCMC Solutions

For the joint distribution (6.4), the full conditional distribution of $\mathbf{z}$ given $\mathbf{x}$ and the parameters is always available as

$$\pi(\mathbf{z}|\mathbf{x}, \theta) \propto \prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i})$$

[6] In non-Bayesian statistics, the EM algorithm is certainly the most ubiquitous numerical method, even though it only applies to (real or artificial) missing variable models.

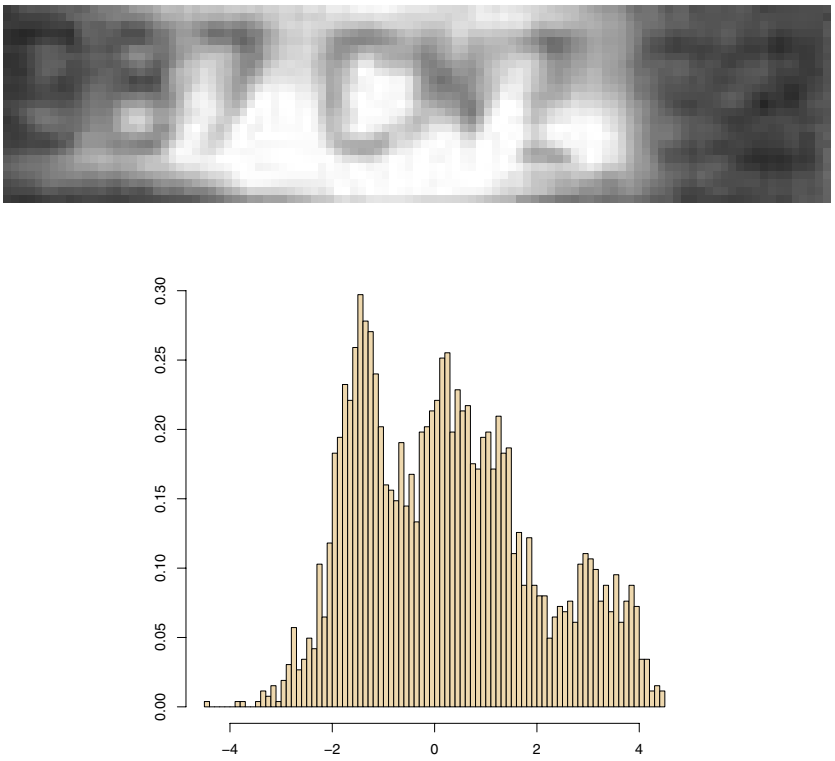


Fig. 6.2. Dataset License: (*top*) Image of a car license plate and (*bottom*) histogram of the transformed grey levels of the dataset.

and can thus be computed at a cost of $O(n)$. Since, for standard distributions $f(\cdot|\theta)$, the full posterior conditionals are also easily simulated when calling on conjugate priors, this implies that the Gibbs sampler can be derived in this setting.[7]

If $\mathbf{p}$ and $\boldsymbol{\theta}$ are independent a priori, then, given $\mathbf{z}$, the vectors $\mathbf{p}$ and $\mathbf{x}$ are independent; that is, $\pi(\mathbf{p}|\mathbf{z},\mathbf{x}) = \pi(\mathbf{p}|\mathbf{z})$. Moreover, in that case, $\boldsymbol{\theta}$ is also independent a posteriori from $\mathbf{p}$ given $\mathbf{z}$ and $\mathbf{x}$, with density $\pi(\boldsymbol{\theta}|\mathbf{z},\mathbf{x})$. If we apply the Gibbs sampler in this problem, it involves the successive simulation of $\mathbf{z}$ and $(\mathbf{p},\boldsymbol{\theta})$ conditional on one another and on the data:

[7]Historically, missing-variable models constituted one of the first instances where the Gibbs sampler was used by completing the missing variables by simulation under the name of *data augmentation* (see Tanner, 1996, and Robert and Casella, 2004, Chapters 9 and 10).

Algorithm 6.1. Mixture Gibbs Sampler
Initialization: Choose $\mathbf{p}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ arbitrarily.
Iteration t $(t \geq 1)$:
1. For $i = 1, \ldots, n$, generate $z_i^{(t)}$ such that

$$\mathbb{P}\left(z_i = j\right) \propto p_j^{(t-1)} f\left(x_i | \theta_j^{(t-1)}\right) .$$

2. Generate $\mathbf{p}^{(t)}$ according to $\pi(\mathbf{p}|\mathbf{z}^{(t)})$.
3. Generate $\boldsymbol{\theta}^{(t)}$ according to $\pi(\boldsymbol{\theta}|\mathbf{z}^{(t)}, \mathbf{x})$.

The simulation of the p_j's is also generally obvious since there exists a conjugate prior (as detailed below). In contrast, the simulation of the θ_j's will strongly depend on the type of sampling density $f(\cdot|\theta)$ as well as the prior π.

The marginal (sampling) distribution of the z_i's is a multinomial distribution $\mathscr{M}_k(p_1, \ldots, p_k)$, which is an exponential family, and thus allows for a conjugate prior, namely the Dirichlet distribution $\mathbf{p} \sim \mathscr{D}\left(\gamma_1, \ldots, \gamma_k\right)$, with density

$$\frac{\Gamma(\gamma_1 + \ldots + \gamma_k)}{\Gamma(\gamma_1) \cdots \Gamma(\gamma_k)} \, p_1^{\gamma_1} \cdots p_k^{\gamma_k}$$

on the simplex of $\mathbb{R}^k$,

$$\mathscr{S}_k = \left\{ (p_1, \ldots, p_k) \in [0,1]^k \, ; \, \sum_{j=1}^{k} p_j = 1 \right\} .$$

In this case, if $n_j = \sum_{l=1}^{n} \mathbb{I}_{z_l = j}$ $(1 \leq j \leq k)$,

$$\mathbf{p}|\mathbf{z} \sim \mathscr{D}\left(n_1 + \gamma_1, \ldots, n_k + \gamma_k\right) .$$

If the density $f(\cdot|\theta)$ also belongs to an exponential family,

$$f(x|\theta) = h(x) \exp\left\{\theta \cdot R(x) - \Psi(\theta)\right\} , \tag{6.6}$$

the simulation of $\boldsymbol{\theta}$ can be specified further. Indeed, we can then use a conjugate prior on each θ_j,

$$\pi(\theta_j) \propto \exp\left\{\theta_j \cdot \xi_j - \lambda_j \Psi(\theta_j)\right\} ,$$

where ξ_j and $\lambda_j > 0$ are hyperparameters. The θ_j's are then independent of one another, given $\mathbf{z}$ and $\mathbf{x}$, with respective distributions

$$\pi(\theta_j|\mathbf{z}, \mathbf{x}) \propto \exp\left\{ \theta_j \cdot \left(\xi_j + \sum_{i=1}^{n} \mathbb{I}_{z_i = j} R(x_i) \right) - \Psi(\theta_j)(\lambda_j + n_j) \right\} ,$$

which are available in closed form, by virtue of the conjugacy.

Exercise 6.5. Show that the θ_j's are dependent on each other given (only) $\mathbf{x}$.

Example 6.3. For the mixture (6.3), under independent normal priors $\mathcal{N}(\delta, 1/\lambda)$ ($\delta \in \mathbb{R}$ and $\lambda > 0$ are fixed hyperparameters) on both μ_1 and μ_2, μ_1 and μ_2 are independent given $(\mathbf{z}, \mathbf{x})$, with conditional distributions

$$\mathcal{N}\left(\frac{\lambda\delta + l\bar{x}_1(\mathbf{z})}{\lambda + \ell}, \frac{1}{\lambda + \ell}\right) \quad \text{and} \quad \mathcal{N}\left(\frac{\lambda\delta + (n-\ell)\bar{x}_2(\mathbf{z})}{\lambda + n - \ell}, \frac{1}{\lambda + n - \ell}\right),$$

respectively. Similarly, the conditional posterior distribution of the z_i's given (μ_1, μ_2) is $(i = 1, \ldots, n)$

$$\mathbb{P}(z_i = 1 | \mu_1, x_i) \propto p \exp\left(-0.5\,(x_i - \mu_1)^2\right).$$

We can thus use the following algorithm to generate from the posterior distribution:

Algorithm 6.2. Mean Mixture Gibbs Sampler

Initialization: Choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$.

Iteration t $(t \geq 1)$:

1. For $i = 1, \ldots, n$, generate $z_i^{(t)}$ from

$$\mathbb{P}(z_i = 1) \propto p \exp\left\{-\frac{1}{2}\left(x_i - \mu_1^{(t-1)}\right)^2\right\},$$

$$\mathbb{P}(z_i = 2) \propto (1-p) \exp\left\{-\frac{1}{2}\left(x_i - \mu_2^{(t-1)}\right)^2\right\}.$$

2. Compute $\ell = \sum_{i=1}^n \mathbb{I}_{z_i^{(t)}=1}$ and $\bar{x}_j(\mathbf{z}) = \sum_{i=1}^n \mathbb{I}_{z_i^{(t)}=j} x_i$.
3. Generate $\mu_1^{(t)}$ from $\mathcal{N}\left(\frac{\lambda\delta + \ell\bar{x}_1(\mathbf{z})}{\lambda + \ell}, \frac{1}{\lambda + \ell}\right)$.
4. Generate $\mu_2^{(t)}$ from $\mathcal{N}\left(\frac{\lambda\delta + (n-\ell)\bar{x}_2(\mathbf{z})}{\lambda + n - \ell}, \frac{1}{\lambda + n - \ell}\right)$.

Figure 6.3 illustrates the usual behavior of this algorithm for the same simulated dataset of 500 points from $0.7\mathcal{N}(0,1) + 0.3\mathcal{N}(2.5,1)$ as in Figure 6.1. The representation of the Gibbs sample over $10,000$ iterations is quite in agreement with the posterior surface, represented here by grey levels and contours, since the second mode discussed in Example 6.1 is much lower than the mode where the simulation output concentrates. ◀

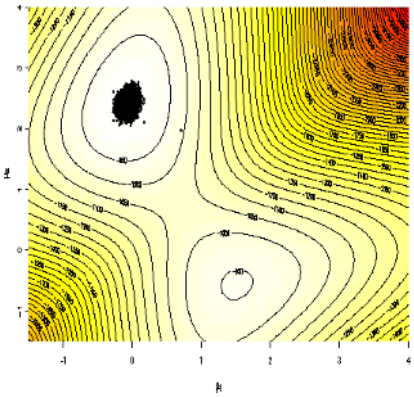


Fig. 6.3. Log-posterior surface and the corresponding Gibbs sample for the model (6.3), based on 10,000 iterations. (The starting point is identifiable on the right of the main mode.)

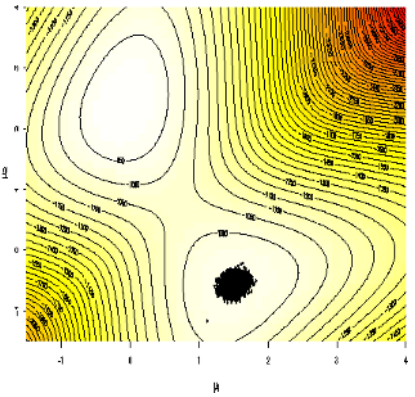


Fig. 6.4. Same graph as Figure 6.3, when initialized close to the second and lower mode, based on 10,000 iterations. (The starting point is identifiable on the left of the second mode.)

⊞ The steps of Algorithm 6.1, extending Algorithm 6.2 to the general normal mixture model, are quite straightforward. For **License** and a normal mixture model, we use $k = 3$ components and we derive the prior distribution from the scale of the problem. Namely, we choose a $\mathscr{D}_3(1/2, 1/2, 1/2)$ prior for the weights (although picking parameters less than 1 in the Dirichlet prior has the potential drawback that it may allow very small weights of some components), a $\mathscr{N}(\overline{x}, \hat{\sigma}^2/3)$ distribution on the means μ_i, and a $\mathscr{G}a(10, \hat{\sigma}^2)$ distribution on the precisions σ_i^{-2}, where $\overline{x}$ and $\hat{\sigma}^2$ are the empirical mean and variance of **License**, respectively. (This empirical choice of a prior is debatable on principle, as it depends on the dataset, but this is relatively harmless since it is equivalent to standardizing the dataset so that the empirical mean and variance are equal to 0 and 1, respectively.) The output represented in Figure 6.5 shows that this crude prior modeling is sufficient to capture the modal features of the histogram as well as the tail behavior in a surprisingly small number of Gibbs iterations, despite the large sample size of 2625 points. The range of the simulated densities represented in Figure 6.5 reflects the variability of the posterior distribution, while the estimate of the density is obtained by averaging the simulated densities over the 500 iterations.[8]

Ⓒ The experiment in Example 6.3 gives a false sense of security about the performance of the Gibbs sampler because it hides the structural dependence of the sampler on its initial conditions. The fundamental feature of Gibbs sampling—its derivation from conditional distributions—implies that it is often restricted in the width of its moves and

[8] That this is a natural estimate of the model, compared with the "plug-in" density using the estimates of the parameters, will be explained more clearly below in Section 6.4.

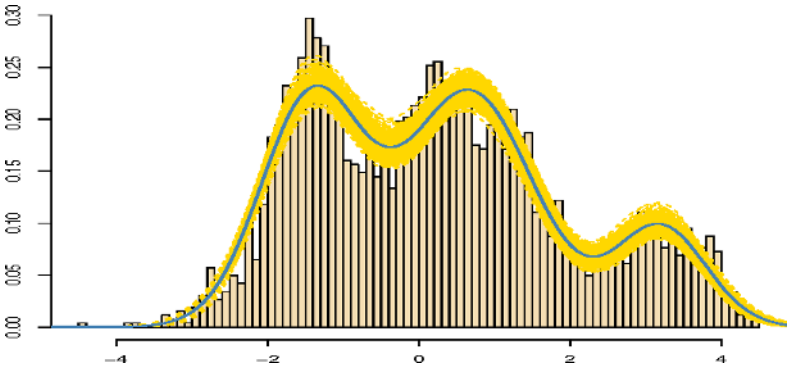


Fig. 6.5. Dataset License: Representation of 500 Gibbs iterations for the mixture estimation. (The accumulated lines correspond to the estimated mixtures at each iteration and the overlaid curve to the density estimate obtained by summation.)

that, in some situations, this restriction may even jeopardize convergence. In the case of mixtures of distributions, conditioning on $\mathbf{z}$ implies that the proposals for $(\boldsymbol{\theta}, \mathbf{p})$ are quite concentrated and do not allow drastic changes in the allocations at the next step. To obtain a significant modification of $\mathbf{z}$ requires a considerable number of iterations once a stable position has been reached.[9] Figure 6.4 illustrates this phenomenon for the same sample as in Figure 6.3: A Gibbs sampler initialized close to the second mode is unable to escape its attraction, even after a large number of iterations, for the reason given above. It is quite interesting to see that this Gibbs sampler suffers from the same pathology as the EM algorithm, although this is not very surprising given that it is based on a similar principle.

In general, there is very little one can do about improving the Gibbs sampler since its components are given by the joint distribution. The solutions are (a) to change the parameterization and thus the conditioning (see Exercise 6.6), (b) to use tempering to facilitate exploration (see Section 6.6), or (c) to mix the Gibbs sampler with another MCMC algorithm.

Exercise 6.6. Construct and test the Gibbs sampler associated with the (ξ, μ_0) parameterization of (6.3), when $\mu_1 = \mu_0 - \xi$ and $\mu_2 = \mu_0 + \xi$.

To look for alternative MCMC algorithms is not a difficulty in this setting, given that the likelihood of mixture models is available in closed form, being computable in $\mathrm{O}(kn)$ time, and the posterior distribution is thus known up to a multiplicative constant. We can therefore use any Metropolis–Hastings

[9] In practice, the Gibbs sampler never leaves the vicinity of a given mode if its attraction is strong enough, as when there are many observations.

algorithm, as long as the proposal distribution q provides a correct exploration of the posterior surface, since the acceptance ratio

$$\frac{\pi(\boldsymbol{\theta}', \mathbf{p}'|\mathbf{x})}{\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x})} \frac{q(\boldsymbol{\theta}, \mathbf{p}|\boldsymbol{\theta}', \mathbf{p}')}{q(\boldsymbol{\theta}', \mathbf{p}'|\boldsymbol{\theta}, \mathbf{p})} \wedge 1$$

can be computed in $\mathrm{O}(kn)$ time. For instance, we can use a random walk Metropolis–Hastings algorithm where each parameter is the mean of the proposal distribution for the new value, that is,

$$\widetilde{\xi}_j = \xi_j^{(t-1)} + u_j,$$

where $u_j \sim \mathscr{N}(0, \zeta^2)$ and ζ is chosen to achieve a reasonable acceptance rate.

For the posterior associated with (6.3), the Gaussian random walk proposal is

$$\widetilde{\mu_1} \sim \mathscr{N}\left(\mu_1^{(t-1)}, \zeta^2\right) \quad \text{and} \quad \widetilde{\mu_2} \sim \mathscr{N}\left(\mu_2^{(t-1)}, \zeta^2\right)$$

associated with the following algorithm:

Algorithm 6.3. Mean Mixture Metropolis–Hastings Sampler

Initialization: Choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$.

Iteration t $(t \geq 1)$:

1. Generate $\widetilde{\mu_1}$ from $\mathscr{N}\left(\mu_1^{(t-1)}, \zeta^2\right)$.
2. Generate $\widetilde{\mu_2}$ from $\mathscr{N}\left(\mu_2^{(t-1)}, \zeta^2\right)$.
3. Compute
$$r = \pi\left(\widetilde{\mu_1}, \widetilde{\mu_2}|x\right) / \pi\left(\mu_1^{(t-1)}, \mu_2^{(t-1)}|x\right)$$
4. Generate $u \sim \mathscr{U}_{[0,1]}$: If $u < r$, then $\left(\mu_1^{(t)}, \mu_2^{(t)}\right) = (\widetilde{\mu_1}, \widetilde{\mu_2})$; otherwise, $\left(\mu_1^{(t)}, \mu_2^{(t)}\right) = \left(\mu_1^{(t-1)}, \mu_2^{(t-1)}\right)$.

For the same simulated dataset as in Figure 6.3, Figure 6.6 shows how quickly this algorithm escapes the attraction of the spurious mode. After a few iterations of the algorithm, the chain drifts away from the poor mode and converges almost deterministically to the proper region of the posterior surface. The Gaussian random walk is scaled as $\zeta = 1$, although slightly smaller scales do work as well but would require more iterations to reach the proper model regions. Too small a scale sees the same trapping phenomenon appear, as the chain does not have sufficient energy to escape the attraction of the current mode (see Example 6.5 and Figure 6.8). Nonetheless, for a large enough scale, the Metropolis–Hastings algorithm overcomes the drawbacks of the Gibbs sampler.

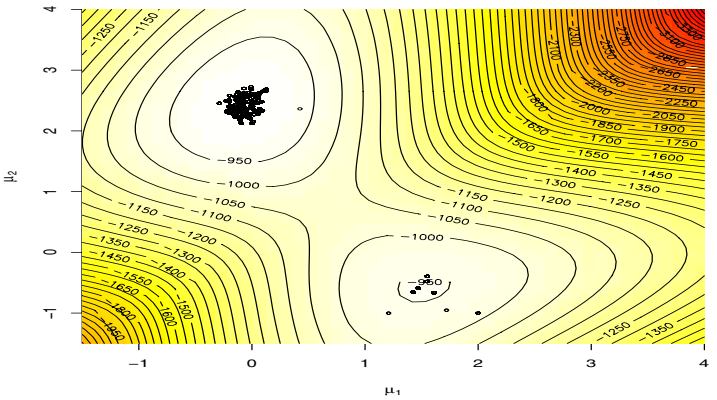


Fig. 6.6. Track of a 10,000 iteration random walk Metropolis–Hastings sample on the posterior surface; the starting point is equal to $(2, -1)$. The scale ζ of the random walk is equal to 1.

We must point out that, for constrained parameters, the random walk Metropolis–Hastings proposal is not efficient because when the chain $(\xi_j^{(t)})$ gets close to the boundary of the parameter space, it moves very slowly, given that the proposed values are often incompatible with the constraint.

Ⓒ This is for instance the case for the weight vector $\mathbf{p}$ since $\sum_{i=1}^k p_k = 1$. A practical resolution of this difficulty is to overparameterize the weights of (6.2)

$$p_j = w_j \Big/ \sum_{l=1}^{k} w_l \,, \quad w_j > 0 \qquad (1 \le j \le k) \,.$$

Obviously, the w_j's are not identifiable, but this is not a difficulty from a simulation point of view and the p_j's remain identifiable (up to a permutation of indices). Perhaps paradoxically, using overparameterized representations often helps with the mixing of the corresponding MCMC algorithms since those algorithms are less constrained by the dataset or by the likelihood. The reader may have noticed that the w_j's are also constrained by a positivity requirement (just like the variances in a normal mixture or the scale parameters for a Gamma mixture), but this weaker constraint can also be lifted using the reparameterization $\eta_j = \log w_j$. The proposed random walk move on the w_j's is thus

$$\log(\widetilde{w_j}) = \log\left\{w_j^{(t-1)}\right\} + u_j \,,$$

where $u_j \sim \mathcal{N}(0, \zeta^2)$. An important difference from the original random walk Metropolis–Hastings algorithm is that the acceptance ratio also involves the proposal via the Jacobian. For instance, the acceptance ratio for a move from $w_j^{(t-1)}$ to $\widetilde{w_j}$ is

$$\frac{\pi(\widetilde{w_j})}{\pi(w_j^{(t-1)})} \, \frac{\widetilde{w_j}}{w_j^{(t-1)}} \wedge 1 \tag{6.7}$$

for this reason.

Exercise 6.7. Give the ratio corresponding to (6.7) when the parameter of interest is in $[0,1]$ and the random walk proposal is on the logit transform $\log\theta/(1-\theta)$.

While being a fairly natural algorithm, the random walk Metropolis–Hastings algorithm usually falls victim to the curse of dimensionality since, obviously the same scale cannot perform well for every component of the parameter vector. In large or even moderate dimensions, a reparameterization of the parameter and preliminary estimation of the information matrix of the distribution are thus most often necessary.

6.4 Label Switching Difficulty

A basic but extremely important feature of a mixture model is that it is invariant under permutations of the indices of the components. For instance, the normal mixtures $0.3\mathcal{N}(0,1)+0.7\mathcal{N}(2.3,1)$ and $0.7\mathcal{N}(2.3,1)+0.3\mathcal{N}(0,1)$ are exactly the same. Therefore, the $\mathcal{N}(2.3,1)$ distribution cannot be called the "first" component of the mixture! In other words, the component parameters θ_i are not identifiable *marginally* in the sense that θ_1 may be 2.3 as well as 0 in the example above. In this specific case, the pairs (θ_1, p) and $(\theta_2, 1-p)$ are exchangeable.

First, in a k-component mixture, the number of modes of the likelihood is of order $\mathrm{O}(k!)$ since if $((\theta_1,\ldots,\theta_k),(p_1,\ldots,p_k))$ is a local maximum of the likelihood function, so is $\tau(\boldsymbol{\theta},\mathbf{p}) = (\theta_{\tau(1)},\ldots,\theta_{\tau(k)})$ for every permutation $\tau \in \mathfrak{S}_k$, the set of all permutations of $\{1,\ldots,k\}$. This makes maximization and even exploration of the posterior surface obviously harder because modes are separated by valleys that most samplers find difficult to cross.

Second, if an exchangeable prior is used on $(\boldsymbol{\theta},\mathbf{p})$ (that is, a prior invariant under permutation of the indices), all the posterior marginals on the θ_i's are identical, which means for instance that the posterior expectation of θ_1 is identical to the posterior expectation of θ_2. Therefore, alternatives to posterior expectations must be considered to provide pertinent estimators.

Example 6.4. In the special case of model (6.3), if we take *the same* normal prior on both μ_1 and μ_2, $\mu_1,\mu_2 \sim \mathcal{N}(0,10)$, say, the posterior weight conditional on $\mathbf{p}$ associated with an allocation $\mathbf{z}$ for which l values are attached to the first component (i.e., such that $\sum_{i=1}^n \mathbb{I}_{z_i=1} = l$) will simply be

$$\omega(\mathbf{z}) \propto p^l(1-p)^{n-l} \int e^{-l(\mu_1-\bar{x}_1)^2/2-(n-1)(\mu_2-\bar{x}_2)^2/2}\, \mathrm{d}\pi(\mu_1)\, \mathrm{d}\pi(\mu_2)$$
$$\times \exp\left(-\left\{s_1^2(\mathbf{z}) - s_2^2(\mathbf{z})\right\}/2\right)$$
$$\propto \sqrt{(l+1/10)(n-l+1/10)}\, p^l(1-p)^{n-l} \exp\left(-\left\{s_1^2(\mathbf{z}) + s_2^2(\mathbf{z})\right.\right.$$
$$\left.\left. + l\bar{x}_1^2/(10l+1) + (n-l)\bar{x}_2^2/(10(n-l)+1)\right\}/2\right),$$

where $s_1^2(\mathbf{z})$ and $s_1^2(\mathbf{z})$ denote the sums of squares for both groups. ◀

Exercise 6.8. Show that, if an exchangeable prior π is used on the vector of weights $(p_1, \ldots, p_k)$, then, necessarily, $\mathbb{E}^\pi[p_j] = 1/k$ and, if the prior on the other parameters $(\theta_1, \ldots, \theta_k)$ is also exchangeable, then $\mathbb{E}^\pi[p_j|x_1, \ldots, x_n] = 1/k$ for all j's.

⊞ For the Gibbs output of **License** discussed above, the exchangeability predicted by the theory is not observed at all, as shown in Figure 6.7. Each component is identified by its mean, and the posterior distributions of the means are very clearly distinct. Although this result has the appeal of providing distinct estimates for the three components, it suffers from the severe drawback, that the Gibbs sampler has not explored the whole parameter space after 500 iterations. Running the algorithm for a much longer period does not solve this problem since the Gibbs sampler cannot simultaneously switch enough component allocations in this highly peaked setup. In other words, the algorithm is unable to explore more than one of the $3! = 6$ equivalent modes of the posterior distribution. Therefore, it is difficult to trust the estimates derived from such an output.

This identifiability problem related to the exchangeability of the posterior distribution, often called "label switching," thus requires either an alternative prior modeling or a more tailored inferential approach. A naïve answer to the problem is to impose an *identifiability constraint* on the parameters, for instance by ordering the means (or the variances or the weights) in a normal mixture (see Exercise 6.3). From a Bayesian point of view, this amounts to truncating the original prior distribution, going from $\pi(\boldsymbol{\theta}, \mathbf{p})$ to

$$\pi(\boldsymbol{\theta}, \mathbf{p})\, \mathbb{I}_{\mu_1 \leq \ldots \leq \mu_k}$$

for instance. While this seems innocuous (given that the sampling distribution is the same with or without this indicator function), the introduction of an identifiability constraint has severe consequences on the resulting inference, both from a prior and from a computational point of view. When reducing the parameter space to its constrained part, the imposed truncation has no reason to respect the topology of either the prior or the likelihood. Instead of singling out one mode of the posterior, the constrained parameter space may then well include parts of several modes and the resulting posterior mean may, for instance, lie in a very low probability region between the modes, while the high posterior probability zones are located at the boundaries of this space.

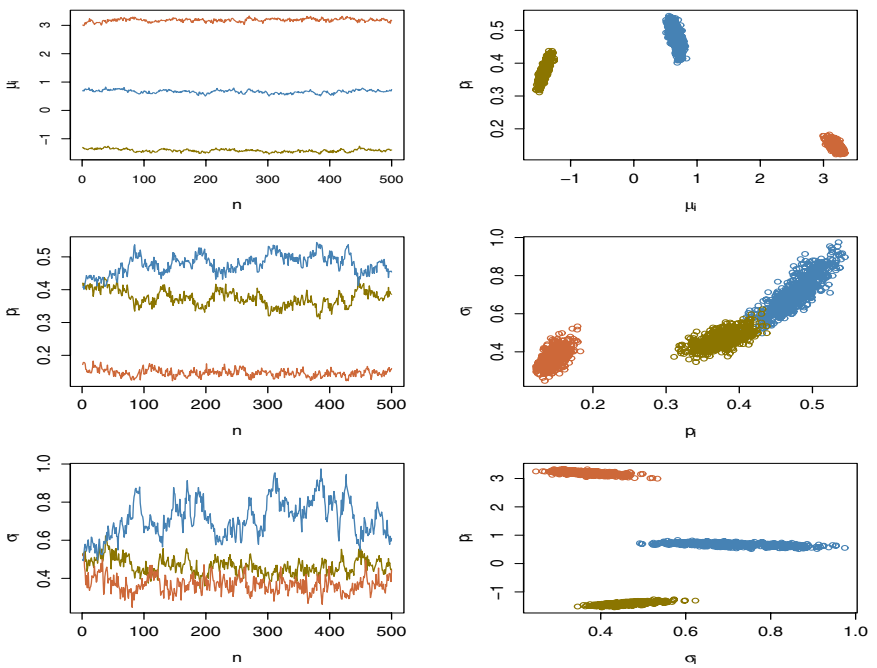


Fig. 6.7. Dataset License: (*left*) Convergence of the three kinds of parameters of the normal mixture; (*right*) 2×2 plot of the Gibbs sample.

In addition, the constraint may radically modify the prior modeling and come close to contradicting the prior information. For large values of k, the introduction of a constraint also has a consequence on posterior inference: With many components, the ordering of components in terms of one of the parameters of the mixture is unrealistic. Some components will be close in mean while others will be close in variance or in weight. This may even lead to very poor estimates of the parameters if the inappropriate ordering is chosen.

Ⓒ While imposing a constraint that is not directly related to the modal regions of the target distribution may considerably reduce the efficiency of an MCMC algorithm, it must be stressed that the constraint does not need to be imposed *during* the simulation but can instead be imposed *after* simulation by reordering the MCMC output according to the constraint. For instance, if the constraint imposes an ordering of the means, once the simulation is over, the components can be relabeled for each MCMC iteration according to this constraint; that is, defining the first component as the one associated with the smallest simulated mean and so on. From this perspective, identifiability constraints have nothing to do with (or against) simulation.

An empirical resolution of the label switching problem that avoids the constraint difficulties altogether consists of selecting one of the $k!$ modal regions

of the posterior distribution once the simulation step is over and only then operate the relabeling in terms of proximity to this region.

Given an MCMC sample of size M, we can find a Monte Carlo approximation of the *maximum a posteriori* (MAP) estimator by taking $\boldsymbol{\theta}^{(i^*)}, \mathbf{p}^{(i^*)}$ such that

$$i^* = \arg \max_{i=1,\ldots,M} \pi \left\{ (\boldsymbol{\theta}, \mathbf{p})^{(i)} | \mathbf{x} \right\} ;$$

that is, the simulated value that gives the maximal posterior density. (Note that π does not need a normalizing constant for this computation.) This value is quite likely to be in the vicinity of one of the $k!$ modes, especially if we run many simulations. The approximate MAP estimate will thus act as a *pivot* in the sense that it gives a good approximation to a mode and we can reorder the other iterations with respect to this mode.

Exercise 6.9. Show that running an MCMC algorithm with target $\pi(\theta|\mathbf{x})^\gamma$ will increase the proximity to the MAP estimate when $\gamma > 1$ is large. (*Note*: This is a crude version of the *simulated annealing* algorithm. See also Chapter 8.) Discuss the modifications required in Algorithm 6.2 to achieve simulation from $\pi(\theta|\mathbf{x})^\gamma$ when $\gamma \in \mathbb{N}^*$.

The selection of the permutations thus reads as follows:

Algorithm 6.4. Pivotal Reordering
At iteration $i \in \{1, \ldots, M\}$:

1. Compute

$$\tau_i = \arg \min_{\tau \in \mathfrak{S}_k} d\left(\tau((\boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i)}), (\boldsymbol{\theta}^{(i^*)}, \mathbf{p}^{(i^*)})) \right),$$

where $d(\cdot, \cdot)$ denotes a distance in the parameter space.

2. Set $(\boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i)}) = \tau_i((\boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i)}))$.

Thanks to this reordering, most iteration labels get switched to the same mode (when n gets large, this is almost a certainty), and the identifiability problem is thus solved. Therefore, after this reordering step, the Monte Carlo estimate of the posterior expectation $\mathbb{E}^\pi_{\mathbf{x}}[\theta_i|\mathbf{x}]$,

$$\sum_{j=1}^{M} (\theta_i)^{(j)} / M \,,$$

can be used as in a standard setting because the reordering automatically gives different meanings to different components. Obviously, $\mathbb{E}^{\pi}_{\mathbf{x}}[\theta_i|\mathbf{x}]$ (or its approximation) should also be compared with $\theta^{(i^*)}$ to check convergence.[10]

6.5 Prior Selection

After insisting in Chapter 2 that conjugate priors are not the only possibility for prior modeling, we seem to be using them quite extensively in this chapter! The fundamental reason for this is that, as explained below, it is not possible to use the standard alternative of noninformative priors on the components. Nonconjugate priors can be used as well (with Metropolis–Hastings steps) but are difficult to fathom when the components have no specific "real" meaning (as, for instance, when the mixture is used as a nonparametric proxy).

The representation (6.2) of a mixture model precludes the use of independent improper priors,

$$\pi\left(\boldsymbol{\theta}\right) = \prod_{i=1}^{k} \pi_i(\theta_i)\,,$$

since if, for any i,

$$\int \pi_i(\theta_i)\mathrm{d}\theta_i = \infty\,,$$

then, for every n,

$$\int \pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{p} = \infty\,.$$

The reason for this inconsistent behavior is that among the k^n terms in the expansion of $\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x})$, there are $(k-1)^n$ terms without *any* observation allocated to the ith component and thus there are $(k-1)^n$ terms with a conditional posterior $\pi(\theta_i|\mathbf{x}, \mathbf{z})$ that is equal to the prior $\pi_i(\theta_i)$.

The inability to use improper priors may be seen by some as a *marginalia*, a fact of little importance, since they argue that proper priors with large variances can be used instead. However, since mixtures are ill-posed problems,[11] this difficulty with improper priors is more of an issue, given that the influence of a particular proper prior, no matter how large its variance, cannot be truly assessed.

[10] While this resolution seems intuitive enough, there is still a lot of debate in academic circles on whether or not label switching should be observed on an MCMC output and, in case it should, on which substitute to the posterior mean should be used.

[11] Ill-posed problems are not precisely defined. They cover classes of models such as *inverse problems*, where the complexity of getting back from the data to the parameters is huge. They are not to be confused with nonidentifiable problems, though.

Prior distributions should always be chosen with the utmost care when dealing with mixtures and their bearing on the resulting inference assessed by a sensitivity study. The fact that some noninformative priors are associated with undefined posteriors, no matter what the sample size, is a clear indicator of the complex nature of Bayesian inference for those models.

Exercise 6.10. In the setting of the mean mixture (6.3), run an MCMC simulation experiment to compare the influence of a $\mathcal{N}(0, 100)$ and of a $\mathcal{N}(0, 10000)$ prior on (μ_1, μ_2) on a sample of 500 observations.

Exercise 6.11. Show that, for a normal mixture $0.5\,\mathcal{N}(0, 1) + 0.5\,\mathcal{N}(\mu, \sigma^2)$, the likelihood is unbounded. Exhibit this feature by plotting the likelihood of a simulated sample using the R image procedure.

6.6 Tempering

The notion of *tempering* can be found in different areas under many different names, but it always comes down to the same intuition that also governs simulated annealing (Chapter 8), namely that when you flatten a posterior surface, it is easier to move around, while if you sharpen it, it gets harder to do so.

More formally, given a density $\pi(x)$, we can define an associated density $\pi_\alpha(x) \propto \pi(x)^\alpha$ for $\alpha > 0$ large enough (if α is too small, $\pi(x)^\alpha$ does not integrate). An important property of this family of distributions is that they all share the same modes. When $\alpha > 1$, the surface of π_α is more contrasted than the surface of π: Peaks are higher and valleys are lower. Increasing α to infinity results in a Dirac mass at the modes of π, and this is the principle behind simulated annealing. Conversely, lowering α to values less than 1 makes the surface smoother by lowering peaks and raising valleys. In a compact space, lowering α to 0 ends up with the uniform distribution.

This rather straightforward remark can be exploited in several directions for simulation. For instance, a tempered version of π, π_α, can be simulated in a preliminary step to determine where the modal regions of π are. (Different values of α can be used in parallel to compare the results.) This preliminary exploration can then be used to build a more appropriate proposal. Alternatively, these simulations may be pursued and associated with appropriate importance weights. Note also that a regular Metropolis–Hastings algorithm may be used with π_α just as well as with π since the acceptance ratio is transformed into

$$\left(\frac{\pi(\boldsymbol{\theta}', \mathbf{p}'|\mathbf{x})}{\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x})}\right)^\alpha \frac{q(\boldsymbol{\theta}, \mathbf{p}|\boldsymbol{\theta}', \mathbf{p}')}{q(\boldsymbol{\theta}', \mathbf{p}'|\boldsymbol{\theta}, \mathbf{p})} \wedge 1 \tag{6.8}$$

in the case of the mixture parameters, with the same independence on the normalizing constants.

Exercise 6.12. Show that the ratio (6.8) goes to 1 when α goes to 0 when the proposal q is a random walk. Describe the average behavior of this ratio in the case of an independent proposal.

Exercise 6.13. If one needs to use importance sampling weights, show that the simultaneous choice of several powers α requires the computation of the normalizing constant of π_α.

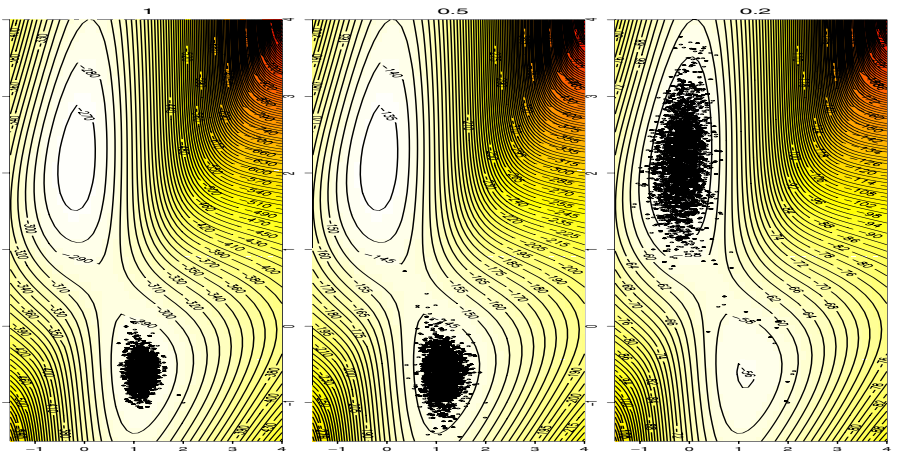


Fig. 6.8. Comparison of Metropolis–Hastings samples of 5000 points started in the vicinity of the spurious mode for the target distributions π_α when $\alpha = 1, 0.5, 0.2$ (*from left to right*), π is the same as in Figure 6.6, and the proposal is a random walk with variance 0.1.

Example 6.5. If we consider once more the posterior associated with (6.3), we can check in Figure 6.8 the cumulative effect of a small variance for the random walk proposal (chosen here as 0.1) and a decrease in the power α. For the genuine target distribution π, 5000 iterations of the Metropolis–Hastings algorithm are not nearly sufficient to remove the attraction of the lower mode. When $\alpha = 0.5$, we can reasonably hope that a few thousand more iterations could bring the Markov chain toward the other mode. For $\alpha = 0.2$, only a few iterations suffice to switch modes, given that the saddle between both modes is not much lower than the modes themselves. (This can be checked by looking at the level sets of the graphs with a good magnifying glass!) ◀

There also exist more advanced techniques that are based on tempering. For instance, we can mention the "pumping mechanism," where a sequence of π_{α_i}'s ($i = 1, \ldots, p$), associated with decreasing α_i's, is used as if one were progressively pressing on the posterior surface (in order to make it flatter) and just as progressively releasing the pressure. The advantage in using this strategy is that (a) it guarantees simulations from the target distribution π and (b) it does not require one to compute or to approximate the normalizing constants of the π_{α_i}'s. The algorithm is as follows, where MCMC(x, π) denotes an arbitrary MCMC step with starting point x and target distribution π:

Algorithm 6.5. Simulated Tempering by Pumping

Initialization: Choose a sequence $1 > \alpha_1 > \cdots > \alpha_p$ and a starting value $x_1^{(1)}$.

iteration t ($t \geq 1$): When in state $x^{(t)}$:

1. Generate $x_1^{(t)}$ from MCMC$(x^{(t)}, \pi_{\alpha_1})$.
2. Generate $x_2^{(t)}$ from MCMC$(x_1^{(t)}, \pi_{\alpha_2})$.

$\vdots$

p. Generate $x_p^{(t)}$ from MCMC$(x_{p-1}^{(t)}, \pi_{\alpha_p})$.

p+1. Generate $x_{p+1}^{(t)}$ from MCMC$(x_p^{(t)}, \pi_{\alpha_{p-1}})$.

$\vdots$

2p–1. Generate $x_{2p}^{(t)}$ from MCMC$(x_{2p-1}^{(t)}, \pi_{\alpha_1})$.

Accept $x^{(t+1)} = x_{2p}^{(t)}$ with probability

$$\min\left\{1, \frac{\pi_{\alpha_1}(x^{(t)})}{\pi(x^{(t)})} \cdots \frac{\pi_{\alpha_p}(x_{p-1}^{(t)})}{\pi_{\alpha_{p-1}}(x_{p-1}^{(t)})} \frac{\pi_{\alpha_{p-1}}(x_p^{(t)})}{\pi_{\alpha_p}(x_p^{(t)})} \cdots \frac{\pi(x_{2p-1}^{(t)})}{\pi^{\alpha_1}(x_{2p-1}^{(t)})}\right\},$$

and take $x^{(t+1)} = x^{(t)}$ otherwise.

Exercise 6.14. Check that this algorithm does not require the normalizing constants of the π_{α_i}'s, and show that π is the corresponding stationary distribution.

The connection with the following section may yet be unclear, but tempering schemes and reversible jump algorithms are related in that the latter are useful for exploring sequences of distributions that usually work on different spaces but may also operate on the same space. Conversely, the purpose of tempering is to facilitate moves around the support of the distribution of interest, and this may involve visiting larger spaces that contain the space of

interest (taking advantage of the intuition that it is sometimes easier to move around obstacles than over them!).

6.7 Variable Dimension Models

While the standard interpretation of mixtures associates a meaning with each component, the nonparametric approach to mixtures only perceives components as base elements in a representation of an unknown density. In that perspective, the number k of components represents the degree of approximation, and it has no particular reason to be fixed in advance. Even from the traditional perspective, it may also happen that the number of homogeneous groups within the population of interest is unknown and that inference first seeks to determine this number. For instance, in a marketing study of Web-browsing behaviors, it may well be that the number of different behaviors is unknown. Alternatively, in the analysis of financial stocks, the number of different patterns in the evolution of these stocks may be unknown to the analyst. For these different situations, it is thus necessary to extend the previous setting to include inference on the number of components itself.

This type of problem belongs to the class of *variable dimension models*. Such models are characterized by a collection of submodels, $\mathfrak{M}_k$, often nested, that are considered simultaneously and associated with different parameter spaces. The number of submodels can be infinite, and the "parameter" is defined conditionally on the index of the submodel, $\boldsymbol{\theta} = (k, \theta_k)$, with a dimension that also depends on k.

Inference on such structures is obviously more complicated than on single models, especially when there are an infinite number of submodels, and it can be tackled from two different (or even opposite) perspectives. The first approach is to consider the variable dimension model as a whole and to estimate quantities that are meaningful for the whole model (such as moments or predictives) as well as quantities that only make sense for submodels (such as posterior probabilities of submodels and posterior moments of θ_k). From a Bayesian perspective, once a prior is defined on $\boldsymbol{\theta}$, the only difficulty is in finding an efficient way to explore the complex parameter space in order to produce these estimators. The second perspective on variable dimension models is to resort to *testing*, rather than estimation, by adopting a *model choice* stance. This requires choosing among all possible submodels the "best one" in terms of an appropriate criterion. The drawbacks of this second approach are far from benign. The computational burden may be overwhelming when the number of models is infinite, the interpretation of the selected model is delicate and the variability of the resulting inference is underestimated since it is impossible to include the effect of the selection of the model in the assessment of the variability of the estimators built in later stages. Nonetheless, this is an

approach often used in linear and generalized linear models (Chapters 3 and 4) where subgroups of covariates are compared against a given dataset.

Mixtures with an unknown number of components are only one particular instance of variable dimension models. As mentioned above, the selection of covariates among k possible covariates in a generalized linear model (Chapter 4) can be seen as a collection of 2^k submodels (depending on the presence or absence of each covariate). Similarly, in a time series model such as the ARMA model (Chapter 7), the value of the lag dependence can be left open, depending on the data at hand. Other instances are the determination of the order in a hidden Markov model (Chapter 7), as in DNA sequences where the dependence of the past bases may go back for one, two, or more steps, or in a capture–recapture experiment when one estimates the number of species from the observed species.

6.7.1 Reversible Jump MCMC

The variable dimension models can be formalized as follows. Take a finite or infinite collection of models $\mathfrak{M}_k$ with corresponding sampling distributions $f_k(\cdot|\theta_k)$ and parameters θ_k, which may live in completely different spaces Θ_k, and associate with each model a prior probability ϱ_k and a prior distribution on the parameter θ_k, $\pi_k(\theta_k)$. Given this formalism, posterior inference can be conducted as usual in the sense that the posterior expectation of a quantity defined across models, such as $\mathbb{E}[x]$, can be computed as

$$\mathbb{E}^{\pi}[\mathbb{E}[x]|x_1,\ldots,x_n] = \sum_k \Pr(\mathfrak{M}_k|x_1,\ldots,x_n) \iint x\, f_k(x|\theta_k)\mathrm{d}x\, \pi_k(\theta_k|x_1,\ldots,x_n)\mathrm{d}\theta\,. \tag{6.9}$$

For quantities that are submodel dependent, it is even simpler since they can be computed directly from $\pi_k(\theta_k|x_1,\ldots,x_n)$. (When a submodel $\mathfrak{M}_k$ is considered separately, like a mixture with k components, we are back to the usual setting of a single model.)

Exercise 6.15. Show that the decomposition (6.9) is correct by representing the generic parameter θ as (k,θ_k) and by introducing the submodel marginals, $m_k(x_1,\ldots,x_n) = \int f_k(x_1,\ldots,x_n|\theta_k)\pi_k(\theta_k)\,\mathrm{d}\theta_k$.

The difficulty in handling this collection of models is in devising an efficient algorithm that explores only submodels that are relevant for the data at hand. When the collection is infinite, this obviously is a requirement: One cannot go through the whole collection by enumeration! Importance sampling techniques are unlikely to work (although they apply in theory) because they need to use

proposals that must simultaneously reach any of the submodels and heavily downweight unlikely submodels.

Exercise 6.16. For a finite collection of submodels $\mathfrak{M}_k$ $(k = 1, \ldots, K)$ with respective priors $\pi_k(\theta_k)$ and weights ϱ_k, write a generic importance-sampling algorithm that approximates the posterior distribution.

While previous algorithms can be used to explore each submodel, they cannot easily bridge the different submodels. For instance, Gibbs sampling cannot be used in this setting, and generic Metropolis–Hastings algorithms suffer from the same drawback as importance sampling in that they need to make move proposals to enough submodels to achieve irreducibility. A specialized MCMC algorithm that meets this efficiency requirement has been proposed by Green (1995). It is called *reversible jump MCMC* (RJMCMC) because of a theoretical requirement of *reversibility* in its validation. While it can be derived from first principles and justified on its own, we will construct Green's algorithm as a special case of the Metropolis–Hastings algorithm.

To explain the specific dynamics of reversible jump MCMC, consider a proposal to move from submodel $\mathfrak{M}_{k_1}$ to submodel $\mathfrak{M}_{k_2}$. (This is not restrictive at all in the sense that, once we are in submodel $\mathfrak{M}_{k_1}$, we need at some point to propose to move to at least another submodel.) A first difficulty is that we are somehow changing target distributions, going from π_{k_1} to π_{k_2} if we condition on the model index k. This is not a serious problem in that the Metropolis–Hastings acceptance ratio extends to this setting (Exercise 6.17). An alternative perspective is that we are always using the same distribution $\pi(k, \theta_k)$ but in different parts of the space.

Exercise 6.17. Show that, if we define the acceptance probability

$$\varrho = \frac{\pi_2(x')}{\pi_1(x)} \frac{q(x|x')}{q(x'|x)} \wedge 1$$

for moving from x to x' and

$$\varrho' = \frac{\pi_1(x)}{\pi_2(x')} \frac{q(x'|x)}{q(x|x')} \wedge 1$$

for the reverse move, the detailed balance condition is modified in such a way that, if $X_t \sim \pi_1(x)$ and if a proposal is made based on $q(x|x_t)$, X_{t+1} is distributed from $\pi_2(x)$. Relate this property to Algorithm 6.5 and its acceptance probability.

The second difficulty is then that we usually have to change dimensions since Θ_{k_1} and Θ_{k_2} may be of different dimensions. The idea at the core

of RJMCMC is to supplement both of the spaces Θ_{k_1} and Θ_{k_2} with adequate artificial spaces in order to create a *one-to-one* mapping between them, most often by augmenting the space of the smaller model. For instance, if $\dim(\Theta_{k_1}) > \dim(\Theta_{k_2})$ and if the move from $\theta_{k_1} \in \Theta_{k_1}$ to Θ_{k_2} is chosen to be a *deterministic* transformation of θ_{k_1},

$$\theta_{k_2} = T_{k_1 \to k_2}(\theta_{k_1}) \,,$$

we can impose a reversibility constraint, namely, that the opposite move from $\theta_{k_2} \in \Theta_{k_2}$ to Θ_{k_1} is concentrated on the curve

$$\mathfrak{K}_{k_1,k_2} = \{\theta_{k_1} \,:\, \theta_{k_2} = T_{k_1 \to k_2}(\theta_{k_1})\} \ .$$

In the general case, θ_{k_1} is completed by a simulation $u_1 \sim g_1(u_1)$ into (θ_{k_1}, u_1) and θ_{k_2} by $u_2 \sim g_2(u_2)$ into (θ_{k_2}, u_2), so that the mapping between (θ_{k_1}, u_1) and (θ_{k_2}, u_2) is a bijection,

$$(\theta_{k_2}, u_2) = T_{k_1 \to k_2}(\theta_{k_1}, u_1), \tag{6.10}$$

while the move from (θ_{k_2}, u_2) to (θ_{k_1}, u_1) is defined by the inverse transform.

If we now apply a *regular* Metropolis–Hastings scheme, proposing a move from the pair (θ_{k_1}, u_1) to (θ_{k_2}, u_2) is like proposing to do so when the corresponding stationary distributions are $\pi(k_1, \theta_{k_1})g_1(u_1)$ and $\pi(k_2, \theta_{k_2})g_2(u_2)$, respectively, and when the proposal distribution is *deterministic* given (6.10). The deterministic feature makes a rather unusual Metropolis–Hastings proposal, but we can solve this difficulty by an approximation. We thus consider instead that the move from (θ_{k_1}, u_1) to (θ_{k_2}, u_2) proceeds by generating

$$(\theta_{k_2}, u_2) \sim \mathscr{N}_{n_2}\left(T_{k_1 \to k_2}(\theta_{k_1}, u_1), \varepsilon I_{n_2}\right) \,, \qquad \varepsilon > 0 \,,$$

where n_2 is the dimension of (θ_{k_2}, u_2), and that the reciprocal proposal is to take (θ_{k_1}, u_1) as the $T_{k_1 \to k_2}$—inverse transform of a normal $\mathscr{N}_{n_2}((\theta_{k_2}, u_2), \varepsilon I_{n_2})$. (This is feasible since $T_{k_1 \to k_2}$ is a bijection.) This reciprocal proposal then has the density

$$\frac{\exp\left\{-\left[T_{k_1 \to k_2}(\theta_{k_1}, u_1) - (\theta_{k_2}, u_2)\right]^2 / 2\varepsilon\right\}}{(2\pi\varepsilon)^{d/2}} \times \left| \frac{\partial T_{k_1 \to k_2}(\theta_{k_1}, u_1)}{\partial(\theta_{k_1}, u_1)} \right|$$

by the Jacobian rule. Therefore, the Metropolis–Hastings acceptance ratio for this *regular* move is

$$1 \wedge \left(\frac{\pi(k_2, \theta_{k_2}) g_2(u_2)}{\pi(k_1, \theta_{k_1}) g_1(u_1)} \left| \frac{\partial T_{k_1 \to k_2}(\theta_{k_1}, u_1)}{\partial(\theta_{k_1}, u_1)} \right| \right.$$
$$\left. \times \frac{\exp\left\{-\left[T_{k_1 \to k_2}(\theta_{k_1}, u_1) - (\theta_{k_2}, u_2)\right]^2 / 2\varepsilon\right\} / (2\pi\varepsilon)^{d/2}}{\exp\left\{-\left[(\theta_{k_2}, u_2) - T_{k_1 \to k_2}(\theta_{k_1}, u_1)\right]^2 / 2\varepsilon\right\} / (2\pi\varepsilon)^{d/2}} \right) \,,$$

and the normal densities cancel as in a regular random walk proposal. If we take into account the fact that several submodels are considered while in models $\mathfrak{M}_1$ and $\mathfrak{M}_2$, with probabilities π_{1i} and π_{2j}, respectively, we end up with the acceptance probability

$$1 \wedge \left(\frac{\pi(k_2,\theta_{k_2})g_2(u_2)\pi_{21}}{\pi(k_1,\theta_{k_1})g_1(u_1)\pi_{12}} \left| \frac{\partial T_{k_1\to k_2}(\theta_{k_1},u_1)}{\partial(\theta_{k_1},u_1)} \right| \right) .$$

As the reader may notice, this probability does not depend on ε. Therefore, we can let ε go to zero and obtain the acceptance probability for the deterministic move (6.10). The acceptance probability for the reverse move is based on the inverse ratio.

The reversible jump algorithm can thus be reinterpreted as a sequence of local fixed-dimensional moves between the models $\mathfrak{M}_k$. The pseudo-code representation of Green's algorithm is thus as follows:

Algorithm 6.6. Green's Reversible Jump Sampler
Iteration t ($t \geq 1$): If $x^{(t)} = (m, \theta^{(m)})$:

1. Select model $\mathfrak{M}_n$ with probability π_{mn}.
2. Generate $u_{mn} \sim \varphi_{mn}(u)$ and set $(\theta^{(n)}, v_{nm}) = T_{m\to n}(\theta^{(m)}, u_{mn})$.
3. Take $x^{(t+1)} = (n, \theta^{(n)})$ with probability

$$\min \left(\frac{\pi(n,\theta^{(n)})}{\pi(m,\theta^{(m)})} \, \frac{\pi_{nm}\varphi_{nm}(v_{nm})}{\pi_{mn}\varphi_{mn}(u_{mn})} \left| \frac{\partial T_{m\to n}(\theta^{(m)},u_{mn})}{\partial(\theta^{(m)},u_{mn})} \right| , 1 \right) , \quad (6.11)$$

and take $x^{(t+1)} = x^{(t)}$ otherwise.

Although the acceptance probability (6.11) is clearly defined, mistakes usually occur either when deriving the Jacobian or when computing the proposal probabilities, as demonstrated in Section 6.7.2.

The presentation above of the reversible jump MCMC algorithm is to be considered as a "soft" introduction to the algorithm and not as a full justification of the method. In particular, this is not at all the path taken by Green (1995). For further theoretical details, the reader may refer to either this seminal paper or to Robert and Casella (2004, Chapter 11).

6.7.2 Reversible Jump for Normal Mixtures

If model $\mathfrak{M}_k$ corresponds to the k-component normal mixture distribution,

$$\sum_{j=1}^{k} p_{jk}\mathcal{N}(\mu_{jk}, \sigma_{jk}^2)\,,$$

we have enormous freedom in choosing the moves from $\mathfrak{M}_k$ to the other models. A simple implementation of Green's algorithm is to restrict the moves from $\mathfrak{M}_k$ only to neighboring models, such as $\mathfrak{M}_{k+1}$ and $\mathfrak{M}_{k-1}$, with probabilities $\pi_{k(k+1)}$ and $\pi_{k(k-1)}$, respectively.

Ⓒ The simplest choice of the corresponding transformation $T_{k\to k+1}$ is a *birth step*, which consists in adding a new normal component in the mixture by generating the parameters of the new component from the prior distribution (which must be proper anyway). The new pair $(\mu_{k+1}, \sigma_{k+1})$ is generated from the corresponding prior on the mean and variance, while the new weight p_{k+1} must be generated from the marginal distribution of p_{k+1} derived from the joint distribution of $(p_1, \ldots, p_{k+1})$ (Exercise 6.18). The Jacobian associated with this move only comes from the renormalization of the weights since the mean and variance are simply added to the current vector of means and variances. In the case of the weights, the previous weights (for the model $\mathfrak{M}_k$) are all multiplied by $(1-p_{k+1})$ to keep the sum equal to 1. The corresponding Jacobian is thus $(1-p_{k+1})^{k-1}$ (given that there are k weights for $\mathfrak{M}_k$ but that one of them is defined as the complement of the sum of the others).

Exercise 6.18. Show that the marginal distribution of p_1 when $(p_1, \ldots, p_k) \sim \mathscr{D}_k(a_1, \ldots, a_k)$ is a $\mathscr{B}e(a_1, a_2 + \ldots + a_k)$ distribution.

The opposite step, the so-called *death step*, is then necessarily derived from the reversibility constraint by removing one of the k components at random. To compute the acceptance probability, we must also take into account the number of possible[12] moves from $\mathfrak{M}_k$ to $\mathfrak{M}_{k+1}$ and from $\mathfrak{M}_{k+1}$ to $\mathfrak{M}_k$. Given a k component mixture, there are $(k+1)!$ ways of defining a $(k+1)$ component mixture by adding one component, while, given a $(k+1)$ component mixture, there are $(k+1)$ choices for the component to die and then $k!$ associated mixtures for the remaining components. We thus end up with a birth acceptance probability equal to

$$\begin{aligned}&\min\left(\frac{\pi_{(k+1)k}}{\pi_{k(k+1)}}\,\frac{(k+1)!}{(k_1)k!}\,\frac{\pi(k+1,\theta_{k+1})}{\pi(k,\theta_k)\,(k+1)\varphi_{k(k+1)}(u_{k(k+1)})}\,,1\right)\\ &\quad=\min\left(\frac{\pi_{(k+1)k}}{\pi_{k(k+1)}}\,\frac{\varrho(k+1)}{\varrho(k)}\,\frac{\ell_{k+1}(\theta_{k+1})\,(1-p_{k+1})^{k-1}}{\ell_k(\theta_k)},1\right)\,,\end{aligned}$$

where ℓ_k denotes the likelihood of the k component mixture model $\mathfrak{M}_k$ and $\varrho(k)$ is the prior probability of model $\mathfrak{M}_k$. Indeed,

$$\pi(k,\theta_k) = \pi(k,\theta_k|\mathbf{x}) \propto \varrho(k)\pi_k(\theta_k)\ell_k(\theta_k)\,,$$

where the normalizing constant *does not* depend on k. The death acceptance probability involves the inverse ratio.

[12] Remember from Section 6.4 that a mixture model is invariant under any of the $k!$ permutations of its components.

This scheme is thus straightforward to implement since we simply make proposals to move the number k of components by plus or minus one and generate the new component parameters from the (marginal) prior distribution if we propose to increase this number by one. Figure 6.9 illustrates this type of proposal starting from a four-component normal mixture.

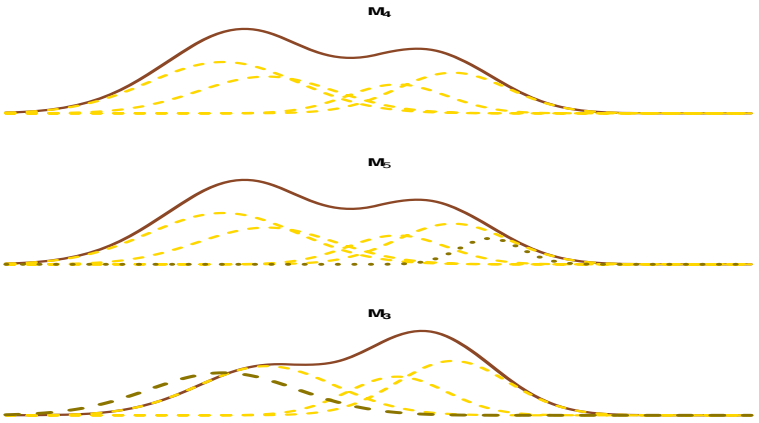


Fig. 6.9. Birth (*middle*) and death (*bottom*) proposals with darker dotted lines, starting from a four-component normal mixture (*top*).

⊞ When applying the birth-and-death scheme to **License**, we manage to get a satisfactory output, even though the mixing is very slow. Using a Poisson $\mathscr{P}oi(4)$ prior on the number of components and the same conjugate priors as above, we thus ran $100,000$ iterations of the RJMCMC and got the output described in Figure 6.10. Out of the $100,000$ proposals to move up or down by one component, only $10,762$ were accepted. Note that, while the density estimate does not change much compared with the fixed $k = 3$ case in Figure 6.5, the number of components and the parameters themselves vary widely. In fact, there is seemingly no stability in either k or the parameter sequences: In Figure 6.10 (*top*), the horizontal lines corresponding to stable locations of the parameters are mostly broken, with new values duplicating previous values. This behavior can be related to the weak identifiability of mixtures that pervades their nonparametric interpretation: Unless small values of the p_i's and σ_i's are excluded by the prior modeling, a given dataset always allows an additional component that hardly modifies the posterior distribution.

While the birth-and-death proposal works well in some settings, it can also be inefficient (that is, leading to a high rejection rate), if the prior is vague, since the birth proposals are not tuned properly (as shown above for **License**). A second proposal, found in Richardson and Green (1997), is to devise more local jumps between models that preserve more structure of the original model. It is made up of *split* and *merge* moves, where a new com-

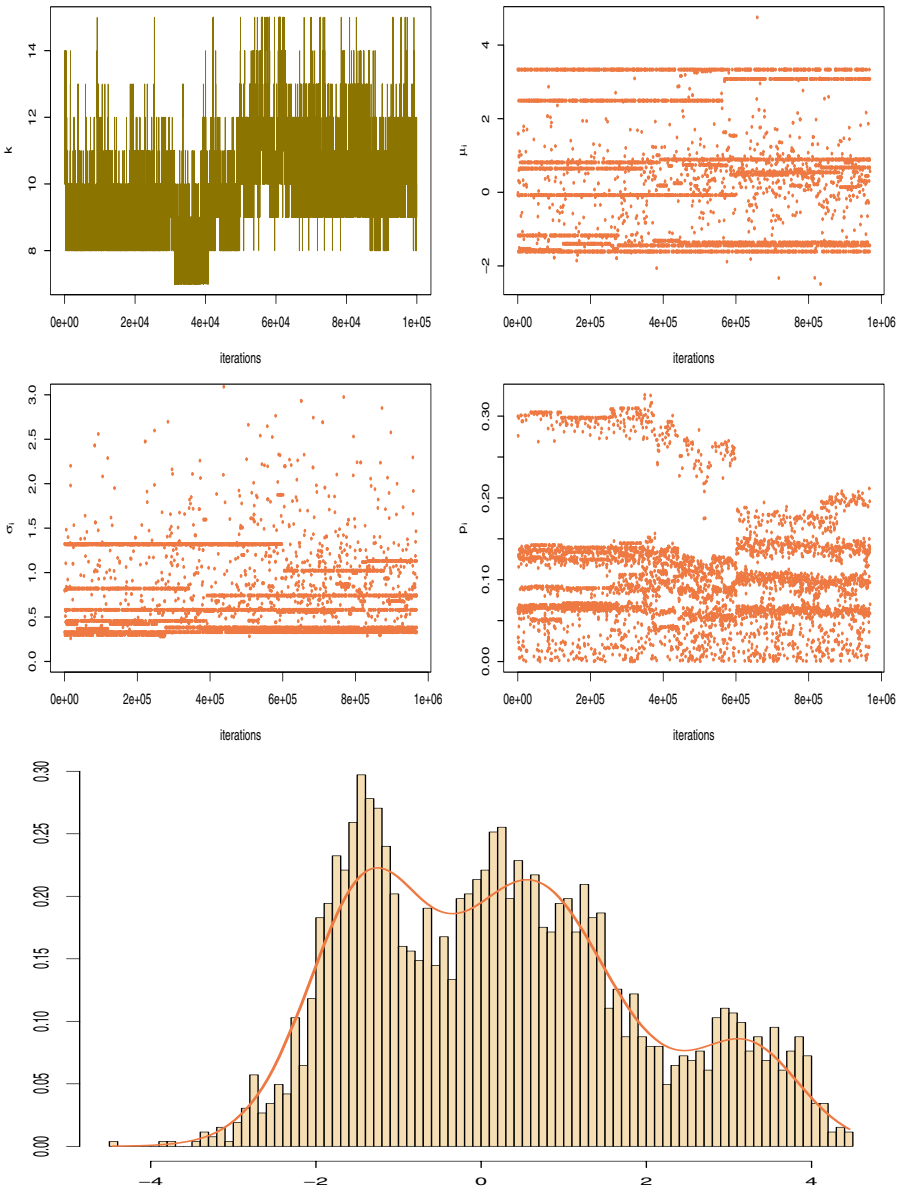


Fig. 6.10. Dataset License: (*top*) Sequences of k, μ_i's, σ_i's, and p_i's in the birth-and-death scheme; (*bottom*) resulting average of the corresponding densities after 100,000 iterations.

ponent is created by splitting an existing component into two, under some moment-preservation conditions, and the reverse move consists of combining two existing components into one, with symmetric constraints that ensure reversibility.

Ⓒ In this split-and-merge proposal, the upward move from $\mathfrak{M}_k$ to $\mathfrak{M}_{k+1}$ replaces a component, say the jth, with two new components, say the jth and the $(j+1)$st, that are *centered* at this earlier component. The split parameters can for instance be created under a moment-preservation condition, which is[13]

$$\begin{aligned} p_{jk} &= p_{j(k+1)} + p_{(j+1)(k+1)}\,, \\ p_{jk}\mu_{jk} &= p_{j(k+1)}\mu_{j(k+1)} + p_{(j+1)(k+1)}\mu_{(j+1)(k+1)}\,, \\ p_{jk}\sigma^2_{jk} &= p_{j(k+1)}\sigma^2_{j(k+1)} + p_{(j+1)(k+1)}\sigma^2_{(j+1)(k+1)}\,. \end{aligned} \tag{6.12}$$

Rather than picking the new component completely at random, we thus force the weight, the mean, and the variance of the selected component to remain constant (in the sense that the mean and the variance of a variable X generated from the new component j with probability $p_{j(k+1)}$ and from the new component $j+1$ with probability $p_{(j+1)(k+1)}$ will be the same as the mean and the variance of a variable X generated from the former component j). The downward move, also called a *merge*, is obtained directly as (6.12) by the reversibility constraint.

A split move satisfying (6.12) can be obtained by generating the auxiliary variable $u_{k(k+1)}$ as $u_1, u_3 \sim \mathcal{U}(0,1)$, $u_2 \sim \mathcal{N}(0,\tau^2)$, and then taking

$$\begin{aligned} p_{j(k+1)} &= u_1 p_{jk}\,, & p_{(j+1)(k+1)} &= (1-u_1)p_{jk}\,, \\ \mu_{j(k+1)} &= \mu_{jk} + u_2\,, & \mu_{(j+1)(k+1)} &= \mu_{jk} - \tfrac{p_{j(k+1)}u_2}{p_{jk}-p_{j)(k+1)}}\,, \\ \sigma^2_{j(k+1)} &= u_3\sigma^2_{jk}\,, & \sigma_{(j+1)(k+1)} &= \tfrac{p_{jk}-p_{j(k+1)}u_3}{p_{jk}-p_{j(k+1)}}\sigma^2_{jk}\,. \end{aligned}$$

If we rank the parameters as $(p_{jk}, u_1, \mu_{jk}, u_2, \sigma^2_{jk}, u_3)$, the corresponding Jacobian of the split transform is thus

$$\begin{pmatrix} u_1 & 1-u_1 & \cdots & \cdots & \cdots & \cdots \\ p_{jk} & -p_{jk} & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & 1 & \cdots & \cdots \\ 0 & 0 & 1 & \frac{-p_{j(k+1)}}{p_{jk}-p_{j(k+1)}} & \cdots & \cdots \\ 0 & 0 & 0 & 0 & u_3 & \frac{p_{jk}-p_{j(k+1)}u_3}{p_{jk}-p_{j(k+1)}} \\ 0 & 0 & 0 & 0 & \sigma^2_{jk} & \frac{-p_{j(k+1)}}{p_{jk}-p_{j(k+1)}}\sigma^2_{jk} \end{pmatrix},$$

with a block diagonal structure that does not require the upper part of the derivatives to be computed. The absolute value of the determinant of this matrix is then

$$p_{jk} \times \frac{p_{jk}}{p_{jk}-p_{j(k+1)}} \times \sigma^2_{jk}\frac{p_{jk}}{p_{jk}-p_{j(k+1)}} = \frac{p_{jk}}{(1-u_1)^2}\,\sigma^2_{jk}\,,$$

and the corresponding acceptance probability is

$$\min\left(\frac{\widetilde{\pi}_{(k+1)k}}{\widetilde{\pi}_{k(k+1)}}\,\frac{\varrho(k+1)}{\varrho(k)}\,\frac{\pi_{k+1}(\theta_{k+1})\ell_{k+1}(\theta_{k+1})}{\pi_k(\theta_k)\ell_k(\theta_k)}\,\frac{p_{jk}}{(1-u_1)^2}\,\sigma^2_{jk}, 1\right),$$

[13] In this case, for clarity's sake, we differentiate the parameters of $\mathfrak{M}_k$ from those of $\mathfrak{M}_{k+1}$ by adding a second index, k or $k+1$, to the component index.

where $\widetilde{\pi}_{(k+1)k}$ and $\widetilde{\pi}_{k(k+1)}$ denote the split and merge probabilities when in models $\mathfrak{M}_k$ and $\mathfrak{M}_{k+1}$, respectively. Once again, the factorial terms vanish: For a split move, there are k possible choices of the split component and then $(k+1)!$ possible orderings of the θ_{k+1} vector, while, for a merge, there are $(k+1)k$ possible choices for the components to be merged and then $k!$ ways of ordering the resulting θ_k.

The reader may wonder how the algorithm can provide efficient approximations to the distribution $\pi(k, \theta_k)$ by continuously changing from one dimension to another. Indeed, it seems as if the algorithm never stays "long enough" in a given submodel! While the validity of this algorithm is as proven as for a fixed-dimension MCMC algorithm, since it is also based on the ergodicity of the resulting chain, and while only moves to relevant values in the next space are accepted, higher efficiency may sometimes be achieved by adding fixed-dimensional moves to the variable-dimensional moves. This hybrid structure is completely valid, with a justification akin to that of the Gibbs sampler; that is, as a composition of several valid MCMC steps. It is even compulsory when some hyperparameters that do not depend on the submodels are used for the prior modeling. Note also that, when using *simultaneously* the birth-and-death and the split-and-merge strategies, we are implicitly using a hybrid MCMC algorithm.

A final point to make is that the missing-data representation is not used in the steps above. If needed (for clustering purposes for instance), the component indicators $\mathbf{z}$ can be simulated as an additional step in the algorithm, but this is not necessary from an algorithmic point of view.

6.7.3 Model Averaging

We can take advantage of this chapter coverage of Bayesian *model choice* to highlight a facet that is uniquely Bayesian, given that it builds a predictive distribution that includes *all* possible models, a feat that is impossible for alternative technologies. Perhaps paradoxically, a natural approach is indeed to integrate over all the models $\mathfrak{M}_k$ as in (6.9). Not only is this a coherent representation from a Bayesian point of view, but it also avoids one of the pitfalls of model choice in that it escapes the usual underestimation of uncertainty resulting from choosing model $\mathfrak{M}_{k_0}$, say, at the model-selection stage, and thereafter ignoring the uncertainty about that choice in the subsequent steps.

This perspective is not appropriate in every setting. A model *must* often be selected, either for scientific (when several theories are compared against a given experiment) or simply monetary reasons (e.g., prohibitive sampling costs). There is also a philosophical objection in favor of parsimony that prohibits keeping complex models when simpler models explain the same data

just as well (this is sometimes called Occam's rule or Occam's razor; see Section 3.5.1 and Robert, 2001, Section 7.4). In addition, as mentioned earlier, this generates an inflation in the number of parameters and, given that most cases involve Monte Carlo or MCMC algorithms, it implies the generation and storage of a vast number of MCMC samples, even though, as mentioned above, MCMC techniques such as reversible jump naturally solve the difficulty of exploring a large number of models by avoiding those with very small probabilities.

In this approach, given a sample $\mathbf{x} = (x_1, \ldots, x_n)$, the predictive distribution is obtained by averaging the model predictives over all models $\mathfrak{M}_k$,

$$\begin{aligned} f(y|\mathbf{x}) &= \int_{\boldsymbol{\Theta}} f(y|\theta)\pi(\theta|\mathbf{x})d\theta \\ &= \sum_k \int_{\Theta_k} f_k(y|\theta_k)\pi(k,\theta_k|\mathbf{x})d\theta_k \\ &= \sum_k p(\mathfrak{M}_k|\mathbf{x}) \int f_k(y|\theta_k)\pi_k(\theta_k|\mathbf{x})\, d\theta_k\,, \end{aligned} \tag{6.13}$$

when $\boldsymbol{\Theta}$ denotes the overall parameter space.

There is a modeling issue at stake in this approach. When facing many alternative models, the choice of the prior probabilities ϱ_k is paramount (as is the choice of the local priors π_k), since

$$p(\mathfrak{M}_k|\mathbf{x}) = \varrho_k m_k(\mathbf{x}) \Big/ \sum_\ell \varrho_\ell m_\ell(\mathbf{x})\,,$$

where m_k denotes the marginal associated with model $\mathfrak{M}_k$, but this choice is also difficult to formalize and justify. For instance, the standard practice of putting equal weights on *all* models is not necessarily a good idea. It does not work if the number of models is infinite and it lacks coherency for nested models (that is, when some models are special cases of others), as in variable selection, since one would think that the bigger model should be more likely than each of its submodels.

In the case of normal mixtures, the use of (6.13) is formally straightforward and results in a predictive distribution that is

$$\sum_{\ell=1}^{K} \mathrm{P}^\pi(k=\ell|\mathbf{x}) \int \sum_{j=1}^{\ell} p_{j\ell} f(y|\theta_{j\ell})\pi_{j\ell}(p_{j\ell},\theta_{j\ell}|\mathbf{x})\, \mathrm{d}p_{j\ell}\, \mathrm{d}\theta_{j\ell}\,, \tag{6.14}$$

where k denotes the unknown number of components and $\pi_{j\ell}$ the marginal posterior distribution on the jth component of the ℓ component mixture.

The label switching phenomenon explained in Section 6.4 seems to strike again in this perspective since, if we choose an exchangeable prior on $(\boldsymbol{\theta}, \mathbf{p})$, all $\pi_{j\ell}$'s are identical (in j, for a fixed ℓ). Therefore, in the event of prior exchangeability, the predictive distribution is written as

$$\sum_{\ell=1}^{K} \mathrm{P}^{\pi}(k=\ell|\mathbf{x}) \int p_{1\ell} f(y|\theta_{1\ell}) \pi_{1\ell}(p_{1\ell}, \theta_{1\ell}|\mathbf{x}) \, \mathrm{d}p_{1\ell} \, \mathrm{d}\theta_{1\ell} \,, \tag{6.15}$$

which appears as a complete cancellation of the mixture structure! However, the fundamental difference with the estimation conundrum (namely that the posterior mean of each pair $(\theta_{j\ell}, p_{j\ell})$ is the same and therefore meaningless) is that the posterior distribution of a *single* component contains the whole information about *all* the components. For instance, in the case of a differentiation of the normal components via their means, the posterior distribution on a mean parameter such as μ_1 will be multimodal, with a mode located near the true mean of each component. Therefore, (6.15) retains the mixture structure, even though it is hidden in $\pi_{1\ell}$. The illustration in the case of the dataset below will make this point more obvious.

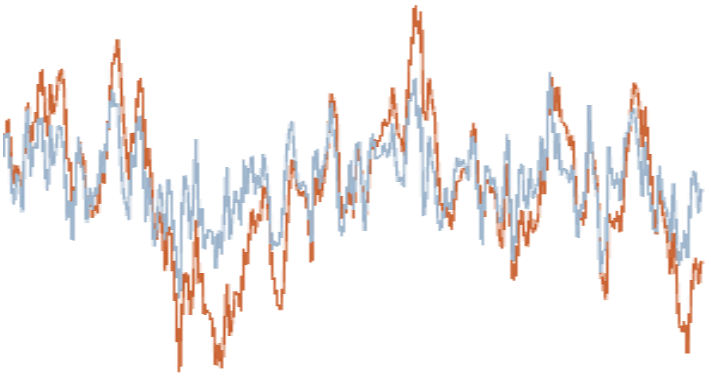
Ⓒ Since the computation of both $\mathrm{P}^{\pi}(k=\ell|\mathbf{x})$ and the $\pi_{j\ell}$'s is impossible in closed form, we must resort to the simulation techniques developed in the previous sections to approximate the predictive. An additional appeal of the model-averaging approach is that it does not require any additional simulation. Given a sequence of T RJMCMC simulations, we can indeed approximate (6.14) by

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{k^{(t)}} p_{jk^{(t)}}^{(t)} f(y|\theta_{jk^{(t)}}^{(t)}) \,,$$

since this approximation converges in T. (The RJMCMC algorithm reproduces through its visits to the different values of k the theoretical frequencies $\mathrm{P}^{\pi}(k=\ell|\mathbf{x})$.) Note that, in the event of posterior exchangeability of the components of the mixture, the Monte Carlo approximation above is unchanged. The fact that it then agrees with both (6.14) *and* (6.15) is due to the fact that the inner sum (over j) above can be interpreted either as in (6.14) as a sum of the components *or* as in (6.15) as a sum over the MCMC iterations (since all component simulations contribute to the approximation of $\pi_{1\ell}$ by virtue of exchangeability). Obviously, the computing cost is not negligible when compared with a usual RJMCMC run since the Monte Carlo approximation must be computed in a series of values of y. However, once a discretization step is chosen on a range of values, the Monte Carlo approximation can be updated at each iteration of the RJMCMC algorithm and does not require enormous storage capacity.

7

Dynamic Models



Rebus was intrigued by the long gaps
in the chronology.
—**Ian Rankin**, ***The Falls***.—

Roadmap

At one point or another, everyone has to face modeling dynamic datasets, by which we mean series of observations that are obviously dependent (like both series in the picture above!). As in the previous chapters, the difficulty in modeling such datasets is to balance the complexity of the representation of the dependence structure against the estimation of the corresponding model—and thus the modeling most often involves model choice or model comparison. We cover here the Bayesian processing of some of the most standard time series models, namely the autoregressive and moving average models, as well as models that are also related to the previous chapter in that the dependence is modeled via a missing-variable structure. These models belong to the category of hidden Markov models and include for instance the stochastic volatility models used in finance. Extended dependence structures met in spatial settings will be briefly considered in Chapter 8. The reader should be aware that, due to special constraints related to the long-term stability of the series, this chapter contains more advanced material.

7.1 Dependent Data

Although independent (iid) observations are the most common objects encountered in statistics textbooks, for the obvious reason that they are easier to process, there is a huge portion of real-life data that does not agree with this independence pattern. As seen in Chapter 5, the capture–recapture model is typical of this disagreement: given the previous capture history (on either the whole population or identified individuals), the captures over time are dependent random variables, as shown by (5.6). This capture–recapture example actually illustrates at once two important cases where dependence may occur, namely *temporal* and *spatial* settings: The capture experiments on the population are done sequentially in time, while the moves of individuals between zones (whether or not these zones are geographically meaningful) induce a spatial pattern.

This chapter concentrates on temporal (or *dynamic*) models, which are somehow simpler because they are unidimensional in their dependence, being indexed only by time. These temporal or *time series* models are some of the most commonly used models in applications, ranging from finance and economics to reliability, to medical experiments, and ecology. This is the case, for instance, for series of pollution data, such as ozone concentration levels, or stock market prices, whose value at time t depends on the previous value at time $t-1$ and also possibly on earlier values.

⊞ Figure 7.1 plots the successive values from January 1, 1998, to November 9, 2003, of the first (in alphabetical order) four stocks[1] of the financial index **Eurostoxx50**, which is a reference for the euro zone[2] constituting 50 major stocks. These values constitute the **Eurostoxx50** dataset. A perusal of these graphs is sufficient for rejecting the assumption of independence of these series: High values are followed by high values and small values by small values, even though the variability (or *volatility*) of the stocks varies from share to share.

The most standard extension of the independence setup is the Markov dependence structure, already mentioned in the simulation techniques of Chapters 5 and 6. We recall that a *stochastic process* $(x_t)_{t\in\mathcal{T}}$ (where, typically, $\mathcal{T}$ is equal to $\mathbb{N}$ or $\mathbb{Z}$) is a Markov chain when the distribution of x_t given the past values (for instance, $\mathbf{x}_{0:(t-1)} = (x_0 \ldots, x_{t-1})$ when $\mathcal{T} = \mathbb{N}$) only depends on

[1] The four stocks are as follows. ABN Amro is an international bank from Holland, with subsidiaries such as Banco Real, LaSalle Bank, or Bank of Asia. Aegon is a Dutch insurance company that focus on life insurance, pensions, savings, and investment products. Ahold Kon., namely Koninklijke Ahold N.V., is also a Dutch company, dealing in retail and food-service businesses, with many subsidiaries in Europe and the United States. Air Liquide is a French company specializing in industrial and medical gases and related services.

[2] At the present time, the euro zone is made up of the following countries: Austria, Belgium, Finland, France, Germany, Greece, Holland, Ireland, Italy, Portugal, and Spain. Other members of the European Union are likely to join in the near future.

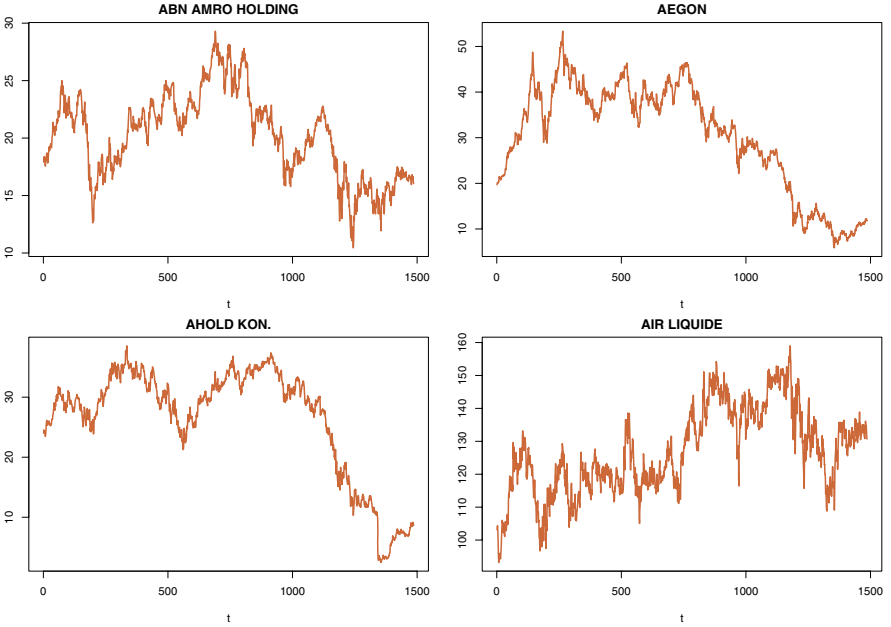


Fig. 7.1. Dataset Eurostoxx50: Evolution of the first four stocks over the period January 1, 1998—November 9, 2003.

x_{t-1}, and this process is *homogeneous* if the distribution of x_t given the past is constant in $t \in \mathcal{T}$. Thus, given an observed sequence $\mathbf{x}_{0:T} = (x_0, \ldots, x_T)$ from a homogeneous Markov chain, the associated likelihood is given by

$$\ell(\theta|\mathbf{x}_{0:T}) = f_0(x_0|\theta) \prod_{t=1}^{T} f(x_t|x_{t-1}, \theta)\,,$$

where f_0 is the distribution of the starting value x_0. From a Bayesian point of view, this likelihood can be processed as in an iid model once a prior distribution on θ is chosen.

This ability to process dynamic models is obviously not limited to the Markov setup. If the likelihood of $\mathbf{x}_{0:T}$, which takes the generic form

$$\ell(\theta|\mathbf{x}_{0:T}) = f_0(x_0) \prod_{t=1}^{T} f_t(x_t|\mathbf{x}_{0:(t-1)}, \theta)\,, \tag{7.1}$$

can be obtained in a closed form, a Bayesian analysis is equally possible.

General models can often be represented as Markov models via the inclusion of missing variables and an increase in the dimension of the model. This is called a *state–space representation*, as described in Section 7.2.3.

While we pointed out above that, once the likelihood function is written down, the Bayesian processing of the model is the same as in the iid case,[3] there exists a major difference that leads to more delicate determination of prior distributions in that *stationarity* and *causality* constraints must often be accounted for. We cannot embark here on a rigorous coverage of stationarity for stochastic processes or even for time series (see Brockwell and Davis, 1996), but we simply mention and motivate the constraints found in the time series literature.

A stochastic process $(x_t)_{t\in\mathcal{T}}$ is stationary[4] if the joint distributions of $(x_1,\ldots,x_k)$ and $(x_{1+h},\ldots,x_{k+h})$ are the same for all indices k and h in $\mathcal{T}$. Formally, this property is called *strict stationarity* because there exists an alternative version of stationarity, called *second–order stationarity*, that imposes the invariance in time only on first and second moments of the process; namely, if we define the autocovariance function $\gamma_x(\cdot,\cdot)$ of $(x_t)_{t\in\mathcal{T}}$ by

$$\gamma_x(r,s) = \mathbb{E}[\{x_r - \mathbb{E}(x_r)\}\{x_s - \mathbb{E}(x_s)\}], \quad r,s\in\mathcal{T}\,,$$

assuming that the variance $\mathbb{V}(x_t)$ is finite, a process $(x_t)_{t\in\mathcal{T}}$ with finite second moments is *second–order stationary* if

$$\mathbb{E}(x_t) = \mu \quad \text{and} \quad \gamma_x(r,s) = \gamma_x(r+t,s+t)$$

for all $r,s,t\in\mathcal{T}$.

If $(x_t)_{t\in\mathcal{T}}$ is second–order stationary, then $\gamma_x(r,s) = \gamma_x(|r-s|,0)$ for all $r,s\in\mathcal{T}$. It is therefore convenient to redefine the autocovariance function of a second–order stationary process as a function of just one variable; i.e., with a slight abuse of notation,

$$\gamma_x(h) = \gamma_x(h,0), \quad h\in\mathcal{T}\,.$$

The function $\gamma_x(\cdot)$ is called the *autocovariance function* of $(x_t)_{t\in\mathcal{T}}$, and $\gamma_x(h)$ is said to be the autocovariance "at lag" h.

Exercise 7.1. Consider the process $(x_t)_{t\in\mathbb{Z}}$ defined by

$$x_t = a + bt + y_t\,,$$

where $(y_t)_{t\in\mathbb{Z}}$ is an iid sequence of random variables with mean 0 and variance σ^2, and where a and b are constants. Define

[3] In the sense that, once a closed form of the posterior is available, there exist generic techniques that do not take into account the dynamic structure of the model.

[4] The connection with the stationarity requirement of MCMC methods is that these methods produce a Markov kernel such that, when the Markov chain is started at time $t=0$ from the target distribution π, the whole sequence $(x_t)_{t\in\mathbb{N}}$ is stationary with marginal distribution π.

$$w_t = (2q+1)^{-1} \sum_{j=-q}^{q} x_{t+j}\,.$$

Compute the mean and the autocovariance function of $(w_t)_{t\in\mathbb{Z}}$. Show that $(w_t)_{t\in\mathbb{Z}}$ is not stationary but that its autocovariance function $\gamma_w(t+h,t)$ does not depend on t.

Obviously, strict stationarity is stronger than second–order stationarity, and it somehow seems more logical.[5] For a process $(x_t)_{t\in\mathbb{N}}$, this obviously relates to the distribution f_0 of the starting values.

Exercise 7.2. Suppose that the process $(x_t)_{t\in\mathbb{N}}$ is such that $x_0 \sim \mathcal{N}(0,\tau^2)$ and, for all $t \in \mathbb{N}$,

$$x_{t+1}|\mathbf{x}_{0:t} \sim \mathcal{N}(x_t/2, \sigma^2)\,, \qquad \sigma > 0\,.$$

Give a necessary condition on τ^2 for $(x_t)_{t\in\mathbb{N}}$ to be a (strictly) stationary process.

Exercise 7.3. Suppose that $(x_t)_{t\in\mathbb{N}}$ is a *Gaussian random walk* on $\mathbb{R}$: $x_0 \sim \mathcal{N}(0,\tau^2)$ and, for all $t \in \mathbb{N}$,

$$x_{t+1}|\mathbf{x}_{0:t} \sim \mathcal{N}(x_t, \sigma^2)\,, \qquad \sigma > 0\,.$$

Show that, whatever the value of τ^2 is, $(x_t)_{t\in\mathbb{N}}$ is not a (strictly) stationary process.

From a Bayesian point of view, to impose the *stationarity* condition on a model (or rather on its parameters) is objectionable on the grounds that the data themselves should indicate whether or not the underlying model is stationary. In addition, since the datasets we consider are always finite, the stationarity requirement is at best artificial in practice. For instance, the series in Figure 7.1 are clearly not stationary on the temporal scale against which they are plotted. However, for reasons ranging from asymptotics (Bayes estimators are not necessarily convergent in nonstationary settings) to causality, to identifiability (see below), and to common practice, it is customary to impose stationarity constraints, possibly on transformed data, even though a Bayesian inference on a nonstationary process could be conducted in principle. The practical difficulty is that, for complex models, the stationarity constraints may get quite involved and may even be unknown in some cases, as for some threshold or changepoint models. We will expose (and solve) this difficulty in the following sections.

Exercise 7.4. Consider the process $(x_t)_{t\in\mathbb{N}}$ such that $x_0 = 0$ and, for all $t \in \mathbb{N}$,

$$x_{t+1}|\mathbf{x}_{0:t} \sim \mathcal{N}(\varrho\, x_t, \sigma^2)\,.$$

[5] Nonetheless, there exists a huge amount of literature on the study of time series based only on second-moment assumptions.

Suppose that $\pi(\varrho, \sigma) = 1/\sigma$ and that there is no constraint on ϱ. Show that the conditional posterior distribution of ϱ, conditional on the observations $\mathbf{x}_{0:T}$ and σ^2, is a $\mathcal{N}(\mu_T, \omega_T^2)$ distribution with

$$\mu_T = \sum_{t=1}^{T} x_{t-1}x_t \Big/ \sum_{t=1}^{T} x_{t-1}^2 \quad \text{and} \quad \omega_T^2 = \sigma^2 \Big/ \sum_{t=1}^{T} x_{t-1}^2 \,.$$

Show that the marginal posterior distribution of ϱ is a Student $\mathcal{T}(T-1, \mu_T, \nu_T^2)$ distribution with

$$\nu_T^2 = \frac{1}{T}\left(\sum_{t=1}^{T} x_t^2 - \sum_{t=1}^{T} x_{t-1}x_t\right) \Big/ \sum_{t=1}^{T} x_{t-1}^2 \,.$$

Apply this modeling to the Aegon series in **Eurostoxx50** and evaluate its predictive abilities.

7.2 Time Series Models

In this section, we consider the most common (linear) time series models, along with their Bayesian analyses and their Markov connections (which can be exploited in MCMC implementations).

7.2.1 AR Models

An AR(1) process $(x_t)_{t\in\mathbb{Z}}$ (where AR stands for autoregressive) is defined by the conditional relation ($t \in \mathbb{Z}$),

$$x_t = \mu + \varrho(x_{t-1} - \mu) + \epsilon_t \,, \tag{7.2}$$

where $(\epsilon_t)_{t\in\mathbb{Z}}$ is an iid sequence of random variables with mean 0 and variance σ^2 (that is, a so-called *white noise*). Unless otherwise specified, we will mostly consider the ϵ_t's to be iid $\mathcal{N}(0, \sigma^2)$ variables.[6]

If $|\varrho| < 1$, $(x_t)_{t\in\mathbb{Z}}$ can be written as

$$x_t = \mu + \sum_{j=0}^{\infty} \varrho^j \epsilon_{t-j} \,, \tag{7.3}$$

and it is easy to see that this is a unique second–order stationary representation. More surprisingly, if $|\varrho| > 1$, the unique second–order stationary representation of (7.2) is

[6] Once again, there exists a statistical approach that leaves the distribution of the ϵ_t's unspecified and only works with first and second moments. But this perspective is clearly inappropriate within the Bayesian framework, which cannot really work with half-specified models.

$$x_t = \mu - \sum_{j=1}^{\infty} \varrho^{-j} \epsilon_{t+j} \,.$$

This stationary solution is frequently criticized as being artificial because it implies that x_t is correlated with the *future* white noises $(\epsilon_t)_{s>t}$, a property not shared by (7.3) when $|\varrho| < 1$. While mathematically correct, the fact that x_t appears as a weighted sum of random variables that are generated after time t is indeed quite peculiar, and it is thus customary to restrict the definition of AR(1) processes to the case $|\varrho| < 1$ so that x_t has a representation in terms of the past realizations $(\epsilon_t)_{s \le t}$. Formally, this restriction corresponds to so-called *causal* or future-independent autoregressive processes.[7] Notice that the causal constraint for the AR(1) model can be naturally associated with a uniform prior on $[-1, 1]$.

If we replace the normal sequence (ϵ_t) with another white noise sequence, it is possible to express an AR(1) process with $|\varrho| > 1$ as an AR(1) process with $|\varrho| < 1$. However, this modification is not helpful from a Bayesian point of view because of the complex distribution of the transformed white noise.

A generalization of the AR(1) model is obtained by increasing the lag dependence on the past values. An AR(p) process is defined by the conditional (against the past) representation ($t \in \mathbb{Z}$),

$$x_t = \mu + \sum_{i=1}^{p} \varrho_i (x_{t+1-i} - \mu) + \epsilon_t \,, \tag{7.4}$$

where $(\epsilon_t)_{t \in \mathbb{Z}}$ is a white noise. As above, we will assume implicitly that the white noise is normally distributed. This natural generalization assumes that the past p values of the process influence (linearly) the current value of the process. Similarly, stationarity and causality constraints can be imposed on this model and, as shown in Brockwell and Davis (1996, Theorem 3.1.1), the AR(p) process (7.4) is both causal and second–order stationary if and only if the roots of the polynomial

$$\mathcal{P}(u) = 1 - \sum_{i=1}^{p} \varrho_i u^i \tag{7.5}$$

are all outside the unit circle in the complex plane. While this necessary and sufficient condition on the parameters ϱ_i is clearly defined, it also imposes an *implicit* constraint on the vector $\boldsymbol{\varrho} = (\varrho_1, \ldots, \varrho_p)$. Indeed, in order to verify that a given vector $\boldsymbol{\varrho}$ satisfies this condition, one needs first to find the roots of the pth degree polynomial $\mathcal{P}$ and then to check that these roots all are of modulus larger than 1. In other words, there is no clearly defined boundary

[7] Both stationary solutions above exclude the case $|\varrho| = 1$. This is because the process (7.2) is then a random walk with no stationary solution.

on the parameter space to define the $\boldsymbol{\varrho}$ that satisfy (or do not satisfy) this constraint, and this creates a major difficulty for simulation applications, given that simulated values of $\boldsymbol{\varrho}$ need to be tested one at a time.

Exercise 7.5. Give necessary and sufficient conditions under which an AR(2) process with autoregressive polynomial $\mathcal{P}(u) = 1 - \varrho_1 u - \varrho_2 u^2$ (with $\varrho_2 \neq 0$) is causal.

The general AR(p) model is Markov, just like the AR(1) model, because the distribution of x_{t+1} only depends on a fixed number of past values. It can thus be expressed as a regular Markov chain when considering the vector, for $t \geq p-1$,

$$\mathbf{z}_t = (x_t, x_{t-1}, \ldots, x_{t+1-p})^{\mathsf{T}} = \mathbf{x}_{t:(t+1-p)}\,.$$

Indeed, we can write

$$\mathbf{z}_{t+1} = \mu \mathbf{1}_p + B(\mathbf{z}_t - \mu \mathbf{1}_p) + \varepsilon_{t+1}\,, \tag{7.6}$$

where

$$\mathbf{1}_p = (1, \ldots, 1)^{\mathsf{T}} \in \mathbb{R}^p\,, \quad B = \begin{pmatrix} \varrho_1 & \varrho_2 & \varrho_3 & \cdots & \varrho_{p-2} & \varrho_{p-1} & \varrho_p \\ 1 & 0 & & \ldots & & & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & & & \ldots & 0 & 1 & 0 \end{pmatrix},$$

and $\varepsilon_t = (\epsilon_t, 0, \ldots, 0)^{\mathsf{T}}$.

If we now consider the likelihood associated with a series $\mathbf{x}_{0:T}$ of observations from an AR(p) process, it depends on the unobserved values $x_{-p}, \ldots, x_{-1}$ since

$$\ell(\mu, \varrho_1, \ldots, \varrho_p, \sigma | \mathbf{x}_{0:T}, \mathbf{x}_{-p:-1}) \propto \sigma^{-T-1} \prod_{t=0}^{T} \exp\left\{ -\left[x_t - \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right]^2 / 2\sigma^2 \right\}.$$

These unobserved initial values can be processed in various ways. First, they can all be set equal to μ, but this is a purely computational convenience with no justification. Second, if the stationarity and causality constraints hold, the process $(x_t)_{t\in\mathbb{Z}}$ has a stationary distribution and one can assume that $\mathbf{x}_{-p:-1}$ is distributed from the corresponding stationary distribution, namely a $\mathcal{N}_p(\mu \mathbf{1}_p, \mathbf{A})$ distribution. We can then integrate those initial values out to obtain the marginal likelihood

$$\int \sigma^{-T-1} \prod_{t=0}^{T} \exp\left\{ \frac{-1}{2\sigma^2} \left(x_t - \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 \right\} f(\mathbf{x}_{-p:-1}|\mu, \mathbf{A})\, \mathrm{d}\mathbf{x}_{-p:-1},$$

based on the argument that they are not directly observed. This likelihood can be dealt with analytically but is more easily processed via a Gibbs sampler that simulates the initial values. An alternative and equally coherent approach is to consider instead the likelihood conditional on the initial *observed* values $\mathbf{x}_{0:(p-1)}$; that is,

$$\ell^c(\mu, \varrho_1, \ldots, \varrho_p, \sigma|\mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)}) \propto \sigma^{-T+p-1} \prod_{t=p}^{T} \exp\left\{ -\left(x_t - \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 / 2\sigma^2 \right\} . \quad (7.7)$$

In this case, if we do not restrict the parameter space through stationarity conditions, a natural conjugate prior can be found for the parameter $\boldsymbol{\theta} = (\mu, \boldsymbol{\varrho}, \sigma^2)$, made up of a normal distribution on $(\mu, \boldsymbol{\varrho})$ and an inverse gamma distribution on σ^2. Instead of the Jeffreys prior, which is controversial in this setting (see Robert, 2001, Note 4.7.2), we can also propose a more traditional noninformative prior such as $\pi(\boldsymbol{\theta}) = 1/\sigma$.

Exercise 7.6. Show that the stationary distribution of $\mathbf{x}_{-p:-1}$ is a $\mathcal{N}_p(\mu \mathbf{1}_p, \mathbf{A})$ distribution, and give an equation satisfied by the covariance matrix $\mathbf{A}$.

Exercise 7.7. Show that the posterior distribution on $\boldsymbol{\theta}$ associated with the prior $\pi(\boldsymbol{\theta}) = 1/\sigma$ is well-defined for $T > p$ observations.

If we do impose the causal stationarity constraint on $\boldsymbol{\varrho}$ that all the roots of $\mathcal{P}$ in (7.5) be outside the unit circle, the set of acceptable $\boldsymbol{\varrho}$'s becomes quite involved and we cannot, for instance, use as prior distribution a normal distribution restricted to this set, if only because we lack a simple algorithm to properly describe the set. While a feasible solution is based on the partial autocorrelations of the AR(p) process (see Robert, 2001, Section 4.5.2), we cover here a different and somehow simpler reparameterization approach using the inverses of the real *and* complex roots of the polynomial $\mathcal{P}$, which are within the unit interval $(-1, 1)$ and the unit sphere, respectively.

If we represent the polynomial (7.5) in its factorized form

$$\mathcal{P}(x) = \prod_{i=1}^{p} (1 - \lambda_i x) ,$$

the inverse roots, λ_i $(i = 1, \ldots, p)$, are either real numbers or complex conjugates.[8] Under the causal stationarity constraint, a natural prior is then to use

[8] The term *conjugate* is to be understood here in the mathematical sense that if $\lambda = \mathfrak{r}\, e^{\iota\theta}$ is a (complex) root of $\mathcal{P}$, then $\bar{\lambda} = \mathfrak{r}\, e^{-\iota\theta}$ is also a (complex) root of $\mathcal{P}$.

uniform priors for these roots, taking a uniform distribution on the number r_p of conjugate complex roots and uniform distributions on $[-1,1]$ and on the unit sphere $\mathscr{S} = \{\lambda \in \mathbb{C};\ |\lambda| \leq 1\}$ for the real and nonconjugate complex roots, respectively. In other words,

$$\pi(\boldsymbol{\lambda}) = \frac{1}{\lfloor p/2 \rfloor + 1} \prod_{\lambda_i \in \mathbb{R}} \frac{1}{2}^{\mathbb{I}_{|\lambda_i|<1}} \prod_{\lambda_i \notin \mathbb{R}} \frac{1}{\pi}^{\mathbb{I}_{|\lambda_i|<1}},$$

where $\lfloor p/2 \rfloor + 1$ is the number of different values of r_p and the second product is restricted to the nonconjugate roots of $\mathcal{P}$. (Note that π is the surface of the unit sphere of $\mathbb{C}$.)

This $\lfloor p/2 \rfloor + 1$ factor, while unimportant for a fixed p setting, must necessarily be included within the posterior distribution when using a reversible jump algorithm to estimate the lag order p since it does not vanish in the acceptance probability of a move between an AR(p) model and an AR(q) model.

While the connection between the inverse roots and the coefficients of the polynomial $\mathcal{P}$ is straightforward (Exercise 7.8), there is no closed-form expression of the posterior distribution either on the roots or the coefficients. Therefore, a numerical approach is once more compulsory to approximate the posterior distribution. However, *any* Metropolis–Hastings scheme can work here, given that the likelihood function can be easily computed in every point.

Exercise 7.8. Show that the coefficients of the polynomial $\mathcal{P}$ can be derived in $\mathsf{O}(p^2)$ time from the inverse roots λ_i using the recurrence relations ($i = 1,\ldots,p, j = 0,\ldots,p$)

$$\psi_0^i = 1\,, \qquad \psi_j^i = \psi_j^{i-1} - \lambda_i \psi_{j-1}^{i-1}\,,$$

where $\psi_0^0 = 1$ and $\psi_j^i = 0$ for $j > i$, and setting $\varrho_j = -\psi_j^p$ ($j = 1,\ldots,p$). Deduce that the likelihood is computable in $\mathsf{O}(Tp^2)$ time.

A rather naïve Metropolis–Hastings implementation that we now describe is to use the prior distribution itself as a proposal on the (inverse) roots of $\mathcal{P}$. This means selecting first one or several roots of $\mathcal{P}$ and then proposing new values for these roots that are simulated from the prior. The acceptance ratio simplifies into the likelihood ratio by virtue of Bayes' theorem. The main difficulty here is that one must take care to modify complex roots by conjugate pairs. This means, for instance, that to create a complex root (and its conjugate) either another complex root (and its conjugate) must be modified or two real roots must be created. We are in a fixed dimensional setting in the sense that, for a fixed value of p, there are a fixed number of parameters but, nonetheless, the real–complex dichotomy is such that the cleanest way to

attack this problem is to use a reversible jump algorithm that distinguishes between the numbers of complex roots.[9]

Ⓒ We thus define the "model" $\mathfrak{M}_{2k}$ $(0 \le k \le \lfloor p/2 \rfloor)$ as corresponding to a number $2k$ of complex roots $(k = 0, \ldots, \lfloor p/2 \rfloor)$, while moving from model $\mathfrak{M}_{2k}$ to model $\mathfrak{M}_{2k+2}$ signifies that two real roots have been replaced by two conjugate complex roots. If we assume that we move from model $\mathfrak{M}_{2k}$ to model $\mathfrak{M}_{2k+2}$ with probability $1/2$ (except when $k = \lfloor p/2 \rfloor$, where the move proposal is always to model $\mathfrak{M}_{2\lfloor p/2 \rfloor - 2}$) and from model $\mathfrak{M}_{2k}$ to model $\mathfrak{M}_{2k-2}$ with probability $1/2$ (except when $k = 0$, where the move proposal is always to model $\mathfrak{M}_2$), the reversible jump[10] acceptance ratio for a move from model $\mathfrak{M}_{2k}$ to model $\mathfrak{M}_{2k+\text{ or }-2}$ is then

$$\frac{\ell^c(\mu, \varrho_1^\star, \ldots, \varrho_p^\star, \sigma | \mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})}{\ell^c(\mu, \varrho_1, \ldots, \varrho_p, \sigma | \mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})} \wedge 1\,, \tag{7.8}$$

except for the extreme cases $k = 0$ and $k = \lfloor p/2 \rfloor$, where the left-hand side of the probability is divided and multiplied by 2, respectively. Note also that the Jacobians cancel in this ratio; that is, the ratio remains the same when the likelihood is expressed in terms of the inverse roots λ_j or in terms of the coefficients ϱ_j.

↯ The determination of the acceptance ratio for a reversible jump algorithm is always a delicate problem, and the utmost care must be taken in its computation.

Ⓒ When using a proper prior, one highly recommended way of checking the validity of the acceptance ratio is to run the algorithm with *no data*. The output should reproduce the prior distribution. Obviously, this is not completely foolproof and does not prevent errors resulting from a wrong computation of the likelihood, for instance, but this helps clean some mistakes in the computation of the move probabilities. For instance, this was how a slight mistake in the code of Richardson and Green (1997) for the estimation of the number of components in a mixture (see Section 6.7.2) was detected. In our case, although we use the improper prior $\pi(\mu, \sigma^2) = 1/\sigma^2$, it is possible to isolate the roots λ_i by fixing σ^2 since they are the only ones involved in the reversible jump moves. For our Algorithm 7.1 below, running the no-data check does provide a uniform distribution on the number of complex roots, as shown by Figures 7.2 and 7.3, and this guarantees that there is no inconsistency in the program. (For instance, there could have been some errors in the choice of the number of ways of choosing complex or real roots and of the number of possible reorderings of those roots.)

[9] A clear advantage of opting for a reversible jump approach is that the extension to the case of the unknown order p is straightforward (Exercise 7.10).

[10] While adopting the variable-dimension perspective is convenient for this problem, it is not imperative to use a reversible jump algorithm. Indeed, building proposals that first generate a number of conjugate complex roots and then generate the vector of roots is also acceptable as long as the moves are efficient.

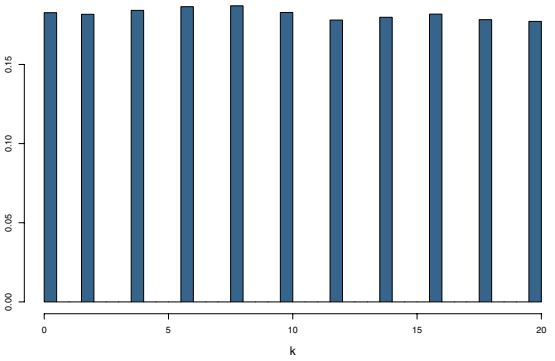


Fig. 7.2. No-data check of Algorithm 7.1 for $p = 20$: This histogram of the number of complex roots shows that the uniform distribution is the stationary distribution and thus validates the corresponding acceptance ratio.

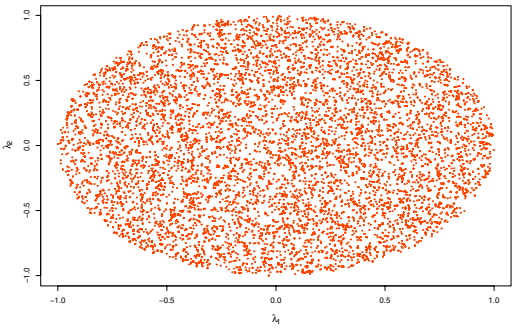


Fig. 7.3. No-data check of Algorithm 7.1 for $p = 20$: This plot of the simulated complex roots shows that the uniform distribution over the unit sphere is well-recovered.

Ⓒ The mean μ and the variance σ^2 being hyperparameters (or, rather, unrelated to the varying dimension problem), they can be updated outside the reversible jump step. Note that it is also possible to add within-steps that modify the parameters λ_i *within a model* $\mathfrak{M}_{2k}$. Once again, if the proposal is the prior distribution, then the acceptance ratio reduces to the likelihood ratio.

Exercise 7.9. Show that, if the proposal on σ^2 is a log-normal distribution $\mathcal{LN}(\sigma^2_{t-1}, \tau^2)$ and if the prior distribution on σ^2 is the noninformative prior $\pi(\sigma^2) = 1/\sigma^2$, the acceptance ratio also reduces to the likelihood ratio because of the Jacobian.

If we consider the conditional likelihood (7.7), which allows a closed-form likelihood (using the transform of Exercise 7.8), a reversible jump algorithm for the AR(p) model and the noninformative prior $\pi(\mu, \sigma) = 1/\sigma$ is as follows:

ALGORITHM 7.1. REVERSIBLE JUMP AR(p) SAMPLER

Initialization: Choose $\boldsymbol{\lambda}^{(0)}$, $\mu^{(0)}$, and $\sigma^{(0)}$.

Iteration t ($t \geq 1$):

1. Select one root at random.
   If the root is real, generate a new real root from the prior distribution. Otherwise, generate a new complex root from the prior distribution and update the conjugate root.
   Replace $\boldsymbol{\lambda}^{(t-1)}$ with $\boldsymbol{\lambda}^\star$ using these new values.
   Calculate the corresponding $\boldsymbol{\varrho}^\star = (\varrho_1^\star, \ldots, \varrho_p^\star)$.
   Take $\boldsymbol{\xi} = \boldsymbol{\lambda}^\star$ with probability
   $$\frac{\ell^c(\mu^{(t-1)}, \boldsymbol{\varrho}^\star, \sigma^{(t-1)}|\mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})}{\ell^c(\mu^{(t-1)}, \boldsymbol{\varrho}^{(t-1)}, \sigma^{(t-1)}|\mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})} \wedge 1\,,$$
   and $\boldsymbol{\xi} = \boldsymbol{\lambda}^{(t-1)}$ otherwise.
2. Select two real roots or two complex conjugate roots at random.
   If the roots are real, generate a new complex root from the prior distribution and compute the conjugate root.
   Otherwise, generate two new real roots from the prior distribution.
   Replace $\boldsymbol{\xi}$ with $\boldsymbol{\lambda}^\star$ using these new values.
   Calculate the corresponding $\boldsymbol{\varrho}^\star = (\varrho_1^\star, \ldots, \varrho_p^\star)$.
   Accept $\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}^\star$ with probability
   $$\frac{\ell^c(\mu^{(t-1)}, \boldsymbol{\varrho}^\star, \sigma^{(t-1)}|\mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})}{\ell^c(\mu^{(t-1)}, \boldsymbol{\varrho}^{(t-1)}, \sigma^{(t-1)}|\mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})} \wedge 1\,,$$
   and set $\boldsymbol{\lambda}^{(t)} = \boldsymbol{\xi}$ otherwise.
3. Generate $\mu^\star$ by a random walk proposal.
   Accept $\mu^{(t)} = \mu^\star$ with probability
   $$\frac{\ell^c(\mu^\star, \boldsymbol{\varrho}^{(t)}, \sigma^{(t-1)}|\mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})}{\ell^c(\mu^{(t-1)}, \boldsymbol{\varrho}^{(t)}, \sigma^{(t-1)}|\mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})} \wedge 1\,,$$
   and set $\mu^{(t)} = \mu^{(t-1)}$ otherwise.
4. Generate $\sigma^\star$ by a log-random walk proposal.
   Accept $\sigma^{(t)} = \sigma^\star$ with probability
   $$\frac{\ell^c(\mu^{(t)}, \boldsymbol{\varrho}^{(t)}, \sigma^\star|\mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})}{\ell^c(\mu^{(t)}, \boldsymbol{\varrho}^{(t)}, \sigma^{(t-1)}|\mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})} \wedge 1\,,$$
   and set $\sigma^{(t)} = \sigma^{(t-1)}$ otherwise.

⊞ As an application of the above, we processed the Ahold Kon. series of **Eurostoxx50**. We ran the algorithm for the whole series with $p = 5$, with satisfactory jump behavior between the different numbers of complex roots. (The same behavior can be observed with larger values of p, but a call to the R ar( ) procedure gives an order of 1 for this series, as

```
> ar(x = Eurostoxx50[, 4])

Coefficients:
     1
0.9968

Order selected 1  sigma^2 estimated as  0.5399
```

This standard analysis is very unstable, being based on the AIC coefficient. For instance, using the alternative

```
>ar(x = Eurostoxx50[, 4], method = "ml")

Coefficients:
     1       2       3      4       5       6       7      8
 1.042  -0.080  -0.038  0.080  -0.049   0.006   0.080  -0.043

Order selected 8  sigma^2 estimated as  0.3228
```

produces a very different order estimate!)

Figure 7.4 summarizes the MCMC output for 5000 iterations. The top left graph shows that jumps between 2 and 0 complex roots occur with high frequency and therefore that the reversible jump algorithm mixes well between both (sub)models. Both following graphs on the first row relate to the hyperparameters μ and σ, which are updated outside the reversible jump steps. The parameter μ appears to be mixing better than σ, which is certainly due to the choice of the same scaling factor in both cases. The middle rows correspond to the first three coefficients of the autoregressive model, $\varrho_1, \varrho_2, \varrho_3$. Their stability is a good indicator of the convergence of the reversible jump algorithm. Note also that, except for ϱ_1, the other coefficients are close to 0 (since their posterior means are approximately 0.052, -0.0001, 2.99×10^{-5}, and -2.66×10^{-7}, respectively). The final row is an assessment of the fit of the model and the convergence of the MCMC algorithm. The first graph provides the sequence of corresponding log-likelihoods, which remain stable almost from the start, the second the distribution of the complex roots, and the last one the connection between the actual series and its one-step-ahead prediction $\mathbb{E}[X_{t+1}|x_t, x_{t-1}, \ldots]$: On this scale, both series are well-related.

Exercise 7.10. Write an R program that extends the reversible jump algorithm above to the case when the order p is unknown and apply it to the same Ahold Kon. series of Eurostoxx50.

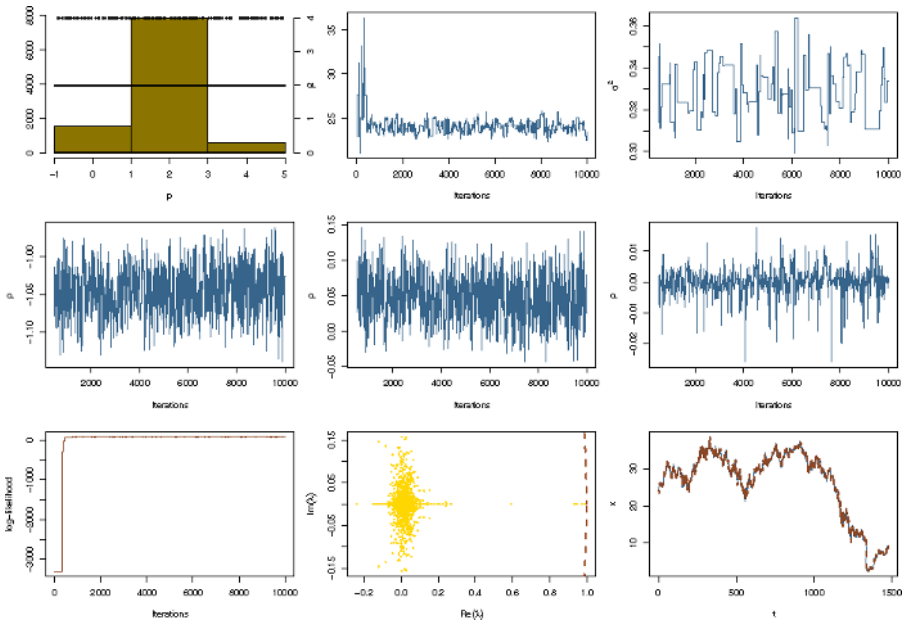


Fig. 7.4. Dataset Eurostoxx50: Output of the reversible jump algorithm for the Ahold Kon. series and an AR(5) model: (*top row, left*) Histogram and sequence of numbers of complex roots (ranging from 0 to 4), (*top row, middle and right*) sequence of μ and σ^2, (*middle row*) sequences of ϱ_i $(i = 1, 2, 3)$, (*bottom row, left*) sequence of observed log-likelihood, (*bottom row, middle*) representation of the cloud of complex roots, with a part of the boundary of the unit circle on the right, (*bottom row, right*) comparison of the series and the one-step-ahead prediction.

7.2.2 MA Models

A second type of time series model that still enjoys linear dependence and closed-form expression is the MA(q) model, where MA stands for moving average. It appears as a dual version of the AR(p) model.

An MA(1) process $(x_t)_{t\in\mathbb{Z}}$ is such that, conditionally on the past $(t \in \mathcal{T})$,

$$x_t = \mu + \epsilon_t - \vartheta\epsilon_{t-1}\,, \tag{7.9}$$

where $(\epsilon_t)_{t\in\mathcal{T}}$ is a white noise sequence. For the same reasons as above, we will assume the white noise is normally distributed unless otherwise specified. Thus,

$$\mathbb{E}[x_t] = \mu\,,\ \mathbb{V}(x_t) = (1+\vartheta^2)\sigma^2\,,\ \gamma_x(1) = \vartheta\sigma^2\,, \quad \text{and} \quad \gamma_x(h) = 0 \quad (h > 1)\,.$$

An important feature of (7.9) is that the model is not identifiable per se. Indeed, we can also rewrite x_t as

$$x_t = \mu + \tilde{\epsilon}_{t-1} - \frac{1}{\vartheta}\tilde{\epsilon}_t, \quad \tilde{\epsilon} \sim \mathcal{N}(0, \vartheta^2\sigma^2)\,.$$

Therefore, both pairs (ϑ, σ) and $(1/\vartheta, \vartheta\sigma)$ are equivalent representations of the *same* model. To achieve identifiability, it is therefore customary in (non-Bayesian environments) to restrict the parameter space of MA(1) processes by $|\vartheta| < 1$, and we will follow suit. Such processes are called invertible. As with causality, the property of invertibility is not a property of the sole process $(x_t)_{t\in\mathbb{Z}}$ but of the connection between both processes $(x_t)_{t\in\mathcal{T}}$ and $(\epsilon_t)_{t\in\mathcal{T}}$.

A natural extension of the MA(1) model is to increase the dependence on the past innovations, namely to introduce the MA(q) process as the process $(x_t)_{t\in\mathcal{T}}$ defined by $(t \in \mathcal{T})$

$$x_t = \mu + \epsilon_t \sum_{i=1}^{q} \vartheta_i \epsilon_{t-i}\,, \tag{7.10}$$

where $(\epsilon_t)_{t\in\mathcal{T}}$ is a white noise (once again assumed to be normal unless otherwise specified). The corresponding identifiability condition in this model is that the roots of the polynomial

$$\mathcal{Q}(u) = 1 - \sum_{i=1}^{q} \vartheta_i u^i$$

are all outside the unit circle in the complex plane (see Brockwell and Davis, 1996, Theorem 3.1.2, for a proof).

The intuition behind the MA(q) representation is less straightforward than the regression structure underlying the AR(p) model. This representation assumes that the dependence between observables stems from a dependence between the (unobserved) noises rather than directly through the observables. Furthermore, in contrast with the AR(p) models, where the covariance between the terms of the series is exponentially decreasing to zero but always different from 0, the autocovariance function for the MA(q) model is such that $\gamma_x(s)$ is equal to 0 for $|s| > q$. In addition, the MA(q) process is obviously (second–order and strictly) *stationary,* whatever the vector $(\vartheta_1, \ldots, \vartheta_q)$, since the white noise is iid and the distribution of (7.10) is thus independent of t. A major difference between the MA(q) and the AR(p) models, though, is that the MA(q) dependence structure is not Markov (even though it can be represented as a Markov process through a state–space representation, introduced below).

While in the Gaussian case the whole observed vector $\mathbf{x}_{1:T}$ is a realization of a normal random variable, with constant mean μ and covariance matrix Σ, and thus provides a formally explicit likelihood function, the computation and obviously the integration (or maximization) of this likelihood are quite costly since they involve inverting the huge matrix Σ.[11]

[11] Obviously, taking advantage of the block diagonal structure of Σ—due to the fact that $\gamma_x(s) = 0$ for $|s| > q$— may reduce the computational cost, but this requires advanced programming abilities!

Exercise 7.11. For an MA(q) process, show that ($s \le q$)

$$\gamma_x(s) = \sigma^2 \sum_{i=0}^{q-|s|} \vartheta_i \vartheta_{i+|s|} \,.$$

A more manageable representation of the MA(q) likelihood is to use the likelihood of $\mathbf{x}_{1:T}$ conditional on the past white noises $\epsilon_0, \ldots, \epsilon_{-q+1}$,

$$\ell^c(\mu, \vartheta_1, \ldots, \vartheta_q, \sigma | \mathbf{x}_{1:T}, \epsilon_0, \ldots, \epsilon_{-q+1}) \propto \sigma^{-T} \prod_{t=1}^{T} \exp\left\{ -\left(x_t - \mu + \sum_{j=1}^{q} \vartheta_j \widehat{\epsilon}_{t-j} \right)^2 / 2\sigma^2 \right\}, \tag{7.11}$$

where ($t > 0$)

$$\widehat{\epsilon}_t = x_t - \mu + \sum_{j=1}^{q} \vartheta_j \widehat{\epsilon}_{t-j} \,,$$

and $\hat{\epsilon}_0 = \epsilon_0$,..., $\hat{\epsilon}_{1-q} = \epsilon_{1-q}$. This recursive definition of the likelihood is still costly since it involves T sums of q terms. Nonetheless, even though the problem of handling the conditioning values $(\epsilon_0, \ldots, \epsilon_{-q+1})$ must be treated separately via an MCMC step, the complexity O(Tq) of this representation is much more manageable than the normal exact representation given above.

Exercise 7.12. Show that the conditional distribution of $(\epsilon_0, \ldots, \epsilon_{-q+1})$ given both $\mathbf{x}_{1:T}$ and the parameters is a normal distribution. Evaluate the complexity of computing the mean and covariance matrix of this distribution.

Ⓒ Given both $\mathbf{x}_{1:T}$ and the past noises $\epsilon_0, \ldots, \epsilon_{-q+1}$, the conditional posterior distribution of the parameters $(\mu, \vartheta_1, \ldots, \vartheta_q, \sigma)$ is formally very close to the posterior associated with an AR(q) posterior distribution. This proximity is such that we can recycle the code of Algorithm 7.1 to some extent since the simulation of the (inverse) roots of the polynomial $\mathcal{Q}$ is identical once we modify the likelihood to the above. The past noises ϵ_{-i} ($i = 1, \ldots, q$) are simulated conditional on the x_t's and on the parameters μ, σ and $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_q)$. While the exact distribution

$$f(\epsilon_0, \ldots, \epsilon_{-q+1} | \mathbf{x}_{1:T}, \mu, \sigma, \boldsymbol{\vartheta}) \propto \prod_{i=-q+1}^{0} e^{-\epsilon_i^2/2\sigma^2} \prod_{t=1}^{T} e^{-\widehat{\epsilon}_t^2/2\sigma^2} \,, \tag{7.12}$$

where the $\widehat{\epsilon}_t$'s are defined as above, is in principle a normal distribution on the vector $(\epsilon_0, \ldots, \epsilon_{-q+1})$ (Exercise 7.12), its computation is too costly to be available for realistic values of T. We therefore implement a hybrid Gibbs algorithm where the missing noise

$\boldsymbol{\epsilon} = (\epsilon_0, \ldots, \epsilon_{-q+1})$ is simulated from a proposal based either on the previous simulated value of $(\epsilon_0, \ldots, \epsilon_{-q+1})$ (in which case we use a simple termwise random walk) or on the first part of (7.12) (in which case we can use normal proposals).[12] More specifically, one can express $\widehat{\epsilon}_t$ $(1 \leq t \leq q)$ in terms of the ϵ_{-t}'s and derive the corresponding (conditional) normal distribution on either each ϵ_{-t} or on the whole vector $\boldsymbol{\epsilon}$[13] (see Exercise 7.13).

ALGORITHM 7.2. REVERSIBLE JUMP MA(q) SAMPLER
Initialization: Choose $\boldsymbol{\lambda}^{(0)}$, $\epsilon^{(0)}$, $\mu^{(0)}$, and $\sigma^{(0)}$ arbitrarily.
Iteration t $(t \geq 1)$:
1. Run steps 1 to 4 of Algorithm 7.1 conditional on $\epsilon^{(t-1)}$ with the correct corresponding conditional likelihood.
2. Simulate $\epsilon^{(t)}$ by a Metropolis–Hastings step.

Exercise 7.13. Give the conditional distribution of ϵ_{-t} given the other ϵ_{-i}'s, $\mathbf{x}_{1:T}$, and the $\widehat{\epsilon}_i$'s. Show that it only depends on the other ϵ_{-i}'s, $\mathbf{x}_{1:q-t+1}$, and $\widehat{\epsilon}_{1:q-t+1}$.

⊞ To illustrate the behavior of this algorithm, we considered the first 350 points of the Air Liquide series in **Eurostoxx50**. The output is represented for $q = 9$ and $10,000$ iterations of Algorithm 7.2, with the same conventions as in Figure 7.4, except that the lower right graph represents the series of the simulated ϵ_{-t} 's rather than the predictive behavior.

Interestingly, the likelihood found by the algorithm as the iteration proceeds is much higher than the one found by the classical R arima procedure since it differs by a factor of 450 on the log scale (assuming we are talking of the same quantity since R arima computes the log-likelihood associated with the observations without the ϵ_{-i}'s!). The details of the call to arima are as follows:

```
> arima(x = Eurostoxx50[1:350, 5], order = c(0, 0, 9))

Coefficients:
         ma1     ma2     ma3     ma4     ma5     ma6     ma7
      1.0605  0.9949  0.9652  0.8542  0.8148  0.7486  0.5574
```

[12] In the following output analysis, we actually used a more hybrid proposal with the innovations $\widehat{\epsilon}_t$'s $(1 \leq t \leq q)$ fixed at their previous values. This approximation remains valid when accounted for in the Metropolis–Hastings acceptance ratio, which requires computing the $\widehat{\epsilon}_t$'s associated with the proposed ϵ_{-i}.

[13] Using the horizon $t = q$ is sensible in this setting given that $x_1, \ldots, x_q$ are the only observations correlated with the ϵ_{-t}'s, even though (7.11) gives the impression of the opposite, given that the $\widehat{\epsilon}_t$'s all depend on the ϵ_{-t}'s.

```
s.e.  0.0531  0.0760  0.0881  0.0930  0.0886  0.0827  0.0774
         ma8     ma9  intercept
      0.3386  0.1300  114.3146
s.e.  0.0664  0.0516  1.1281

sigma^2 estimated as 8.15:  log likelihood = -864.97
```

The favored number of complex roots is 6, and the smaller values 0 and 2 are not visited after the initial warmup. The mixing over the σ parameter is again lower than over the mean μ, despite the use of three different proposals. The first one is based on the inverted gamma distribution associated with $\widehat{\epsilon}_{-(q-1):q}$, the second one is based on a (log) random walk with scale $0.1\hat{\sigma}_x$, and the third one is an independent inverted gamma distribution with scale $\hat{\sigma}_x/(1+\vartheta_1^2+\ldots+\vartheta_q^2)^{1/2}$. Note also that, except for ϑ_9, the other coefficients ϑ_i are quite different from 0 (since their posterior means are approximately 1.0206, 0.8403, 0.8149, 0.6869, 0.6969, 0.5693, 0.2889, and 0.0895, respectively). This is also the case for the estimates above obtained in R arima. The prediction being of little interest for MA models (Exercise 7.14), we represent instead the range of simulated ϵ_t's in the bottom right figure. The range is compatible with the $\mathcal{N}(0,\sigma^2)$ distribution.

Exercise 7.14. Show that the predictive horizon for the MA(q) model is restricted to the first q future observations x_{t+i}.

7.2.3 State–Space Representation of Time Series Models

An alternative approach of considerable interest for the representation and analysis of the MA(q) model and its generalizations is the so-called *state–space representation*, which relies on missing variables to recover both the Markov structure *and* the linear framework.[14]

The general idea is to represent a time series $(\mathbf{x}_t)$ as a system of two equations,

$$\mathbf{x}_t = G\mathbf{y}_t + \varepsilon_t\,, \tag{7.13}$$

$$\mathbf{y}_{t+1} = F\mathbf{y}_t + \xi_t\,, \tag{7.14}$$

where ε_t and ξ_t are multivariate normal vectors[15] with general covariance matrices that may depend on t and $\mathbb{E}[\varepsilon_u^{\mathsf{T}}\xi_v] = 0$ for all (u,v)'s. Equation (7.13) is called the *observation equation*, while (7.14) is called the *state equation*. This representation embeds the process of interest $(\mathbf{x}_t)$ into a larger space, the *state*

[14] It is also inspired from the *Kalman filter*, ubiquitous for prediction, smoothing, and filtering in time series.

[15] Notice the different fonts that distinguish the ε_t's used in the state–space representation from the ϵ_t's used in the AR and MA models.

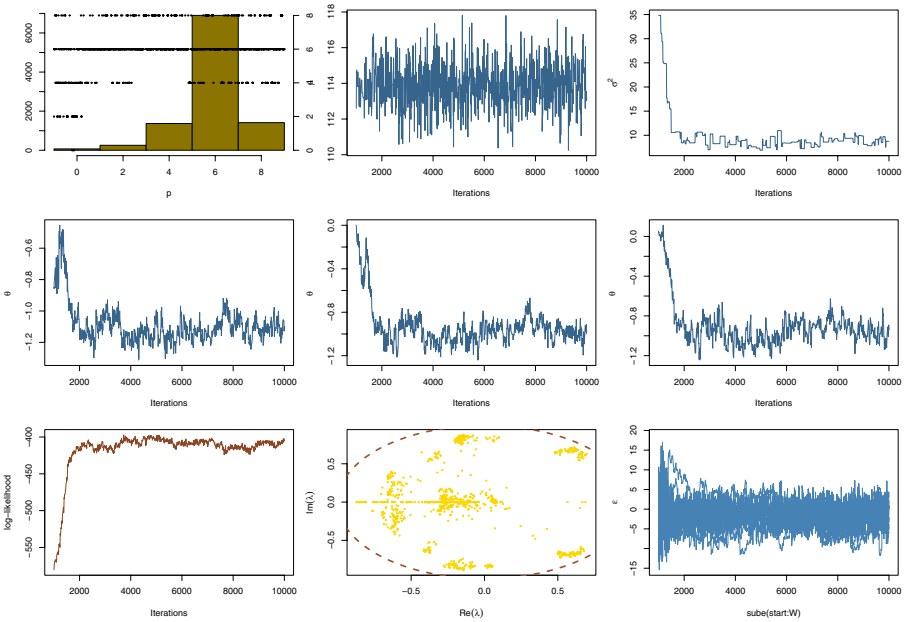


Fig. 7.5. Dataset Eurostoxx50: Output of the reversible jump algorithm for the Air Liquide series and an MA(9) model: (*top row, left*) Histogram and sequence of numbers of complex roots (ranging from 0 to 8); (*top row, middle and right*) sequence of μ and σ^2; (*middle row*) sequences of ϑ_i $(i = 1, 2, 3)$; (*bottom row, left*) sequence of observed likelihood; (*bottom row, middle*) representation of the cloud of complex roots, with the boundary of the unit circle; and (*bottom row, right*) evolution of the simulated ϵ_{-t} 's.

space, where the *missing* process $(\mathbf{y}_t)$ is Markov and linear. For instance, (7.6) is a *state–space representation* of the AR(p) model.

The MA(q) model can be written that way by defining $\mathbf{y}_t$ as

$$(\epsilon_{t-q}, \ldots, \epsilon_{t-1}, \epsilon_t)^{\mathsf{T}} .$$

Then the state equation is

$$\mathbf{y}_{t+1} = \begin{pmatrix} 0\,1\,0\,\ldots\,0 \\ 0\,0\,1\,\ldots\,0 \\ \qquad\ldots \\ 0\,0\,0\,\ldots\,1 \\ 0\,0\,0\,\ldots\,0 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} , \tag{7.15}$$

while the observation equation is

$$\mathbf{x}_t = x_t = \mu - \left(\vartheta_q\ \vartheta_{q-1}\ \ldots\ \vartheta_1\ -1\right) \mathbf{y}_t \,,$$

with no perturbation ε_t.

The state–space decomposition of the MA(q) model thus involves no vector ε_t in the observation equation, while ξ_t is degenerate in the state equation. The degeneracy phenomenon is quite common in state–space representations, but this is not a hindrance in conditional uses of the model, as in MCMC implementations. Notice also that the state–space representation of a model is not unique, again a harmless feature for MCMC uses. For instance, for the MA(1) model, the observation equation can also be chosen as $x_t = \mu + (1 \;\; 0)\mathbf{y}_t$ with $\mathbf{y}_t = (y_{1t}, y_{2t})^{\mathsf{T}}$ directed by the state equation

$$\mathbf{y}_{t+1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \mathbf{y}_t + \begin{pmatrix} 1 \\ -\vartheta_1 \end{pmatrix} \epsilon_{t+1}\,.$$

7.2.4 ARMA Models

A straightforward extension of both previous models are the (normal) ARMA(p, q) models, where x_t ($t \in \mathbb{Z}$) is conditionally defined by

$$x_t = \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) + \epsilon_t - \sum_{j=1}^{q} \vartheta_j \epsilon_{t-j}\,, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)\,, \tag{7.16}$$

the (ϵ_t)'s being independent. The role of such models, as compared with both AR and MA models, is to aim toward parsimony; that is, to use much smaller values of p and q than in a pure AR(p) or a pure MA(q) modeling.

The causality and invertibility conditions on the parameters of (7.16) still correspond to the roots of both polynomials $\mathcal{P}$ and $\mathcal{Q}$ being outside the unit circle, respectively, with a further condition that both polynomials have no common root. (But this almost surely never happens under a continuous prior on the parameters.) The root reparameterization can therefore be implemented for both the ϑ_i's and the ϱ_j's, still calling for MCMC techniques owing to the complexity of the posterior distribution.

State–space representations also exist for ARMA(p, q) models, one possibility being

$$\mathbf{x}_t = x_t = \mu - \left(\vartheta_{r-1}\; \vartheta_{r-2} \ldots \vartheta_1 \; -1\right) \mathbf{y}_t$$

for the observation equation and

$$\mathbf{y}_{t+1} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & & \ldots & \\ 0 & 0 & 0 & \ldots & 1 \\ \varrho_r & \varrho_{r-1} & \varrho_{r-2} & \ldots & \varrho_1 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \tag{7.17}$$

for the state equation, with $r = \max(p, q+1)$ and the convention that $\varrho_m = 0$ if $m > p$ and $\vartheta_m = 0$ if $m > q$.

Similarly to the MA(q) case, this state–space representation is handy in devising MCMC algorithms that converge to the posterior distribution of the parameters of the ARMA(p, q) model.

Ⓒ A straightforward MCMC processing of the ARMA model is to take advantage of the AR and MA algorithms that have been constructed above by using both algorithms sequentially. Indeed, conditionally on the AR parameters, the ARMA model can be expressed as an MA model and, conversely, conditionally on the MA parameters, the ARMA model can be expressed almost as an AR model. This is quite obvious for the MA part since, if we define ($t > p$)

$$\tilde{x}_t = x_t - \mu + \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu)\,,$$

the likelihood is formally equal to a standard MA(q) likelihood on the $\tilde{x}_t$'s. The reconstitution of the AR(p) likelihood is more involved: If we define the residuals $\tilde{\epsilon}_t = \sum_{j=1}^{q} \vartheta_j \epsilon_{t-j}$, the log-likelihood conditional on $\mathbf{x_{0:(p-1)}}$ is

$$-\sum_{t=p}^{T} \left(x_t - \mu - \sum_{j=1}^{p} \varrho_j [x_{t-j} - \mu] - \tilde{\epsilon}_t \right)^2 \Big/ 2\sigma^2\,,$$

which is obviously close to an AR(p) log-likelihood, except for the $\tilde{\epsilon}_t$'s. The original AR(p) code can then be recycled modulo this modification in the likelihood.

Another extension of the AR model is the ARCH model, used to represent processes, particularly in finance, with independent errors with time-dependent variances, as in the ARCH(p) process[16] ($t \in \mathbb{Z}$)

$$x_t = \sigma_t \epsilon_t\,, \qquad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)\,, \qquad \sigma_t^2 = \alpha + \sum_{i=1}^{p} \beta_i x_{t-i}^2\,.$$

The ARCH(p) process defines a Markov chain since x_t only depends on $\mathbf{x}_{t-p:t-1}$. It can be shown that a stationarity condition for the ARCH(1) model is that $\mathbb{E}[\log(\beta_1 \epsilon_t^2)] < 0$, which is equivalent to $\beta_1 < 3.4$. This condition becomes much more involved for larger values of p. Contrary to the stochastic volatility model defined in Example 7.1 below, the ARCH(p) model enjoys a closed-form likelihood when conditioning on the initial values $x_1, \ldots, x_p$. However, because of the nonlinearities in the variance terms, approximate methods based on MCMC algorithms must be used for their analysis.

7.3 Hidden Markov Models

Hidden Markov models are simultaneously a generalization of the mixture models of Chapter 6 and a generalization of state–space models. Their appeal

[16] The acronym ARCH stands for *autoregressive conditional heteroscedasticity*, heteroscedasticity being a term favored by econometricians to describe heterogeneous variances. Gouriéroux (1996) provides a general reference on these models, as well as classical inferential methods of estimation.

within this chapter is that they constitute an interesting case of non-Markov time series besides being extremely useful in modeling financial, telecommunication, and genetic data. We refer the reader to McDonald and Zucchini (1997) for a deeper introduction to these models and to Cappé et al. (2004) for complete coverage of their statistical processing.

7.3.1 Basics

The family of *hidden Markov models* (abbreviated to HMM) consists of a bivariate process $(x_t, y_t)_{t\in\mathbb{N}}$, where the subprocess $(y_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain on a state space $\mathscr{Y}$ and, conditional on $(y_t)_{t\in\mathbb{N}}$, $(x_t)_{t\in\mathbb{N}}$ is a series of random variables on $\mathscr{X}$ such that the conditional distribution of x_t only depends on y_t, as represented by the DAG in Figure 7.6. When $\mathscr{Y} = \{1, \ldots, \kappa\}$, we have, in particular,

$$x_t|y_t \sim f(x|\xi_{y_t})$$

where $(y_t)_{t\in\mathbb{N}}$ is a finite state–space Markov chain (i.e., $y_t|y_{t-1}$ is distributed from $\mathbb{P}(y_t = i|y_{t-1} = j) = p_{ji}$ $(1 \le i \le \kappa)$) and the ξ_i's are the different values of the parameter indexing the conditional distribution. In the general case, the joint distribution of (x_t, y_t) given the past values $\mathbf{x}_{0:(t-1)} = (x_0, \ldots, x_{t-1})$ and $\mathbf{y}_{0:(t-1)} = (y_0, \ldots, y_{t-1})$ factorizes as

$$(x_t, y_t)|\mathbf{x}_{0:(t-1)}, \mathbf{y}_{0:(t-1)} \sim f(y_t|y_{t-1})\, f(x_t|y_t)\,,$$

in agreement with Figure 7.6. The process $(y_t)_{t\in\mathbb{N}}$ is usually referred to as the *state* of the model and is *not observable* (hence, *hidden*). Inference thus has to be carried out only in terms of the observable process $(x_t)_{t\in\mathbb{N}}$.

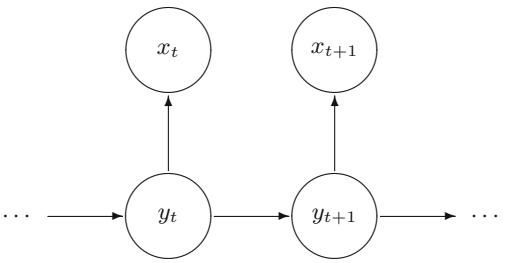


Fig. 7.6. Directed acyclic graph (DAG) representation of the dependence structure of a hidden Markov model, where $(x_t)_{t\in\mathbb{N}}$ is the observable process and $(y_t)_{t\in\mathbb{N}}$ the hidden process.

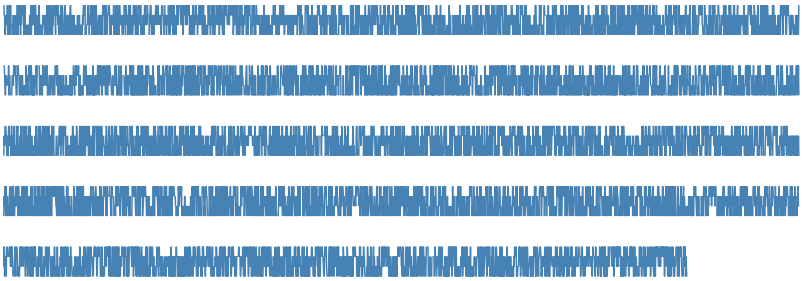


Fig. 7.7. Dataset Dnadataset: Sequence of 9718 amine bases for an HIV genome. The four bases A, C, G, and T have been recoded as $1, ..., 4$.

Exercise 7.15. Show that, when the support $\mathcal{Y}$ is finite and when $(y_t)_{t\in\mathbb{N}}$ is stationary, the marginal distribution of x_t is the same mixture distribution for all t's. Deduce that the same identifiability problem as in mixture models occurs in this setting.

⊞ Hidden Markov models have been used in genetics since the early 1990s for the modeling of DNA sequences. In short, DNA, which stands for deoxyribonucleic acid, is a molecule that carries the genetic information about a living organism and is replicated in each of its cells. This molecule is made up of a sequence of amine bases—adenine, cytosine, guanine, and thymine—abbreviated as A, C, G, and T. The particular arrangement of bases in different parts of the sequence is thought to be related to different characteristics of the living organism to which it corresponds. **Dnadataset** is a particular sequence corresponding to a complete HIV (which stands for Human Immunodeficiency Virus) genome where A, C, G, and T have been recoded as $1, ..., 4$. Figure 7.7 represents this sequence of 9718 bases. The simplest modeling of this sequence is to assume a two-state hidden Markov model with $\mathscr{Y} = \{1, 2\}$ and $\mathscr{X} = \{1, 2, 3, 4\}$, the assumption being that one state corresponds to noncoding regions and the other to coding regions.

State–space models are thus a special case of hidden Markov models in the sense that the equations (7.13) and (7.14) are a special occurrence of the generic representation

$$\begin{aligned} \mathbf{x}_t &= G(\mathbf{y}_t, \epsilon_t)\,, \\ \mathbf{y}_t &= F(\mathbf{y}_{t-1}, \zeta_t)\,. \end{aligned} \tag{7.18}$$

Note, however, that it is not necessarily appealing to use this representation, in comparison with state–space models, because the complexity of the functions F or G may hinder the processing of this representation to unbearable levels (while, for state–space models, the linearity of the relations always allows generic processing like Gibbs sampling steps).

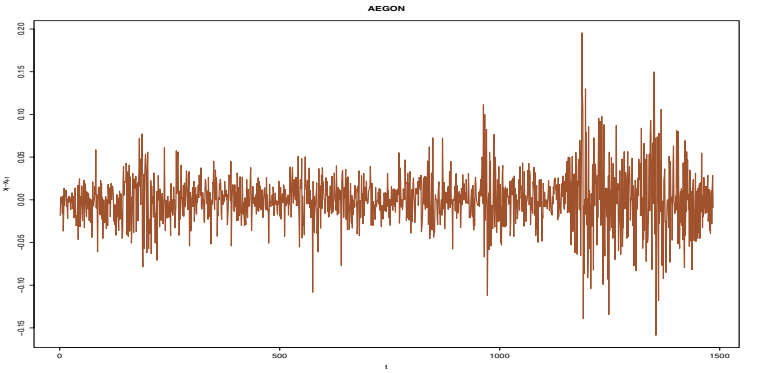


Fig. 7.8. Dataset Eurostoxx50: First-order difference $\{\log(x_t) - \log(x_{t-1})\}$ of the Aegon stock sequence regarded as a potential stochastic volatility process (7.19).

Example 7.1. Stochastic volatility models are quite popular in financial applications, especially in describing series with sudden and correlated changes in the magnitude of variation of the observed values. These models use a hidden chain $(y_t)_{t\in\mathbb{N}}$, called the *stochastic volatility*, to model the variance of the observables $(x_t)_{t\in\mathbb{N}}$ in the following way: Let $y_0 \sim \mathcal{N}(0, \sigma^2)$ and, for $t = 1, \ldots, T$, define

$$\begin{cases} y_t = \varphi y_{t-1} + \sigma\epsilon^*_{t-1}\,, \\ x_t = \beta e^{y_t/2}\epsilon_t\,, \end{cases} \tag{7.19}$$

where both ϵ_t and ϵ^*_t are iid $\mathcal{N}(0,1)$ random variables. In this simple version, the observable is thus a white noise, except that the variance of this noise enjoys a particular AR(1) structure on the logarithmic scale. Quite obviously, this structure makes the computation of the (observed) likelihood a formidable challenge! ◀

⊞ Figure 7.8 gives the sequence $\{\log(x_t) - \log(x_{t-1})\}$ when (x_t) is the Aegon stock sequence plotted in Figure 7.1. While this real-life sequence is not necessarily a stochastic volatility process, it presents some features that are common with those processes, including an overall stationary structure and periods in the magnitude of the variation of the sequence.

Exercise 7.16. Write down the joint distribution of $(y_t, x_t)_{t\in\mathbb{N}}$ in (7.19) and deduce that the (observed) likelihood is not available in closed form.

As in Example 7.1, the distributions of both x_t and y_t are usually parameterized, that is, (7.18) looks like

$$\begin{aligned} \mathbf{x}_t &= \ G(\mathbf{y}_t, \epsilon_t|\theta)\,, \\ \mathbf{y}_t &= F(\mathbf{y}_{t-1}, \zeta_t|\delta)\,, \end{aligned} \tag{7.20}$$

where θ and δ are finite-dimensional parameters.

To draw inference on either the parameters of the HMM or on the hidden chain (as is again the case for the stochastic volatility model in Example 7.1), it is generally necessary to take advantage of the missing-variable nature of HMMs and to use simultaneous simulation both of $(y_t)_{t\in\mathbb{N}}$ and the parameters of the model. There is, however, one exception to that requirement, which is revealed in Section 7.3.2, and that is when the state space $\mathcal{Y}$ of the hidden chain $(y_t)_{t\in\mathbb{N}}$ is finite.

In the event that both the hidden and the observed chains are finite, with $\mathcal{Y} = \{1, \ldots, \kappa\}$ and $\mathcal{X} = \{1, \ldots, k\}$, as in Dnadataset, the parameter θ is made up of p probability vectors $\mathbf{q}^1 = (q_1^1, \ldots, q_k^1), \ldots, \mathbf{q}^\kappa = (q_1^\kappa, \ldots, q_k^\kappa)$ and the parameter δ is the $\kappa \times \kappa$ Markov transition matrix $\mathbb{P} = (p_{ij})$ on $\mathcal{Y}$. Given that the joint distribution of $(x_t, y_t)_{0\le t\le T}$ is

$$\varrho_{y_0}\, q_{x_0}^{y_0} \prod_{t=1}^{T} p_{y_{t-1}y_t}\, q_{x_t}^{y_t}\,,$$

where $\boldsymbol{\varrho} = (\varrho_1, \ldots, \varrho_\kappa)$ is the stationary distribution of $\mathbb{P}$ (i.e., such that $\boldsymbol{\varrho}\mathbb{P} = \boldsymbol{\varrho}$), the posterior distribution of (θ, δ) given $(x_t, y_t)_t$ factorizes as

$$\pi(\theta,\delta)\, \varrho_{y_0} \prod_{i=1}^{\kappa}\prod_{j=1}^{k} (q_j^i)^{n_{ij}} \times \prod_{i=1}^{\kappa}\prod_{j=1}^{p} p_{ij}^{m_{ij}}\,,$$

where the n_{ij}'s and the m_{ij}'s are sufficient statistics representing the number of visits to state j by the x_t's when the corresponding y_t's are equal to i and the number of transitions from state i to state j on the hidden chain $(y_t)_{t\in\mathbb{N}}$, respectively. If we condition on the starting value y_0 (and thus omit ϱ_{y_0} in the expression above) and if we use a flat prior on the p_{ij}'s and q_j^i's, the posterior distributions are Dirichlet. Similarly to the ARMA case, if we include the starting values in the posterior distribution, this introduces a non-conjugate structure in the simulation of the p_{ij}'s, but this can be handled with a Metropolis–Hastings substitute that uses the Dirichlet distribution as the proposal. Note that, in any case, we need to simulate y_0.

When comparing HMMs with ARMA models, it may appear that the latter are more general in the sense that they allow a different dependence on the past values. Resorting to the state–space representation (7.18) shows that this is not the case. Different horizons p of dependence can also be included for HMMS simply by (a) using a vector $\mathbf{x}_t = (x_{t-p+1}, \ldots, x_t)$ for the observables or by (b) using a vector $\mathbf{y}_t = (y_{t-q+1}, \ldots, y_t)$ for the latent process in (7.18).

© Once the parameters are simulated, the simulation of the chain $(y_t)_{0\le t\le T}$ can be processed Gibbs-wise (i.e., one term at a time), using the fully conditional distributions

$$\mathbb{P}(y_t = i | x_t, y_{t-1}, y_{t+1}) \propto p_{y_{t-1}i}\, p_{iy_{t+1}}\, q_{x_t}^i\,.$$

Algorithm 7.3. Finite–State HMM Gibbs Sampler

Initialization:

1. Generate random values of the p_{ij}'s and the q_j^i's.
2. Generate the hidden Markov chain $(y_t)_{0\le t\le T}$ by $(i = 1, 2)$

$$\mathbb{P}(y_t = i) \propto \begin{cases} p_{ii}\, q^i_{x_0} & \text{if } t = 0\,, \\ p_{y_{t-1}i}\, q^i_{x_t} & \text{if } t > 0\,, \end{cases}$$

and compute the corresponding sufficient statistics.

Iteration m $(m \ge 1)$:

1. Generate

$$(p_{i1}, \ldots, p_{i\kappa}) \sim \mathscr{D}(1 + n_{i1}, \ldots, 1 + n_{i\kappa})\,,$$
$$(q^i_1, \ldots, q^i_k) \sim \mathscr{D}(1 + m_{i1}, \ldots, 1 + m_{ik})\,,$$

and correct for the missing initial probability by a Metropolis–Hastings step with acceptance probability $\varrho'_{y_0}/\varrho_{y_0}$.

2. Generate successively each y_t $(0 \le t \le T)$ by

$$\mathbb{P}(y_t = i | x_t, y_{t-1}, y_{t+1}) \propto \begin{cases} p_{ii}\, q^i_{x_1}\, p_{iy_1} & \text{if } t = 0\,, \\ p_{y_{t-1}i}\, q^i_{x_t}\, p_{iy_{t+1}} & \text{if } t > 0\,, \end{cases}$$

and compute the corresponding sufficient statistics.

In the initialization step of Algorithm 7.3, any distribution on $(y_t)_{t\in\mathbb{N}}$ is obviously valid, but this particular choice is of interest since it is related to the true conditional distribution, simply omitting the dependence on the next value.

⊞ For **Dnadataset**, the dimensions are $\kappa = 2$ and $k = 4$. We ran several Gibbs samplers for 1000 iterations, starting from small, medium and high values for p_{11} and p_{22}, and got very similar results in both first and both last cases for the approximations to the Bayes posterior means, as shown by Table 7.1. The raw output also gives a sense of stability, as shown by Figure 7.9.

For the third case, started at small values of both p_{11} and p_{22}, the simulated chain had not visited the same region of the posterior distribution after those 1000 iterations, and it produced an estimate with a smaller log-likelihood[17] value of $-13,160$. However, running the Gibbs sampler longer (for 4000 more iterations) did produce a similar estimate, as shown by the third replication in Table 7.1. This phenomenon is slightly related to the phenomenon, discussed in the context of Figures 6.3 and 6.4, that the Gibbs sampler tends to "stick" to lower modes for lack of sufficient energy. In the current situation, the energy required to leave the lower mode appears to be available. Note that we have reordered the output to compensate for a possible switch between hidden states 1 and 2 among experiments. This is quite natural, given the lack of identifiability

[17] The log-posterior is proportional to the log-likelihood in that special case, and the log-likelihood was computed using a technique described in Section 7.3.2.

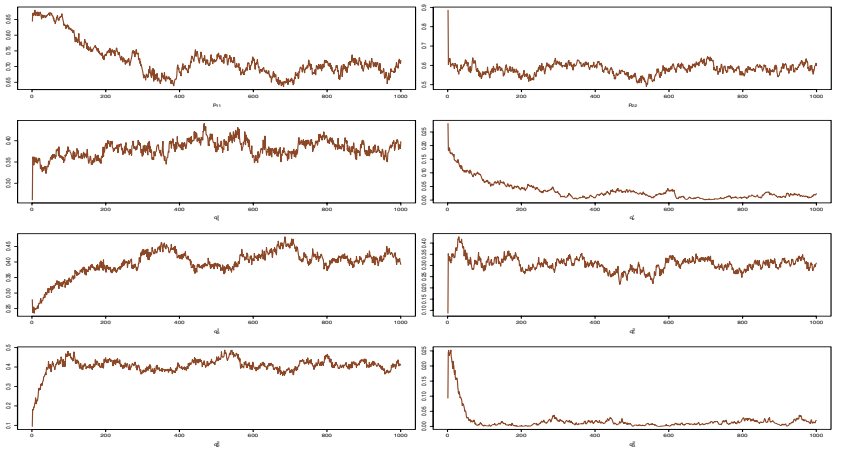


Fig. 7.9. Dataset Dnadataset: Convergence of a Gibbs sequence to the region of interest on the posterior surface for the hidden Markov model (this is replication 2 in Table 7.1). The row-wise order of the parameters is the same as in Table 7.1.

of the hidden states (Exercise 7.15). Flipping the indices 1 and 2 does not modify the likelihood, and thus all these experiments explore the same mode of the posterior.

Table 7.1. Dataset Dnadataset: Five runs of the Gibbs sampling approximations to the Bayes estimates of the parameters for the hidden Markov model along with final log-likelihood (starting values are indicated on the line below in parentheses) based on $M = 1000$ iterations (except for replication 3, based on 5000 iterations).

Repl.	p_{11}	p_{22}	q_1^1	q_2^1	q_3^1	q_1^2	q_2^2	q_3^2	Log-like.
1	0.720	0.581	0.381	0.032	0.396	0.306	0.406	0.018	–13,121
	(0.844)	(0.885)	(0.260)	(0.281)	(0.279)	(0.087)	(0.094)	(0.0937)	
2	0.662	0.620	0.374	0.016	0.423	0.317	0.381	0.034	–13,123
	(0.628)	(0.621)	(0.203)	(0.352)	(0.199)	(0.066)	(0.114)	(0.0645)	
3	0.696	0.609	0.376	0.023	0.401	0.318	0.389	0.030	–13,118
	(0.055)	(0.150)	(0.293)	(0.200)	(0.232)	(0.150)	(0.102)	(0.119)	
4	0.704	0.580	0.377	0.024	0.407	0.313	0.403	0.020	–13,121
	(0.915)	(0.610)	(0.237)	(0.219)	(0.228)	(0.079)	(0.073)	(0.076)	
5	0.694	0.585	0.376	0.0218	0.410	0.315	0.395	0.0245	–13,119
	(0.600)	(0.516)	(0.296)	(0.255)	(0.288)	(0.110)	(0.095)	(0.107)	

7.3.2 Forward–Backward Representation

When the state space of the hidden Markov chain $\mathcal{Y}$ is finite, that is, when

$$\mathcal{Y} = \{1, \ldots, \kappa\} \ ,$$

the likelihood function[18] of the observed process $(x_t)_{1\le t\le T}$ can be computed in manageable $\mathrm{O}(T\times\kappa^2)$ time by a recurrence relation called the *forward–backward* or *Baum–Welch* formulas.[19]

As illustrated in Figure 7.6, a generic feature of HMMs is that ($t = 2,\ldots,T$)

$$p(y_t|y_{t-1},\mathbf{x}_{0:T}) = p(y_t|y_{t-1},\mathbf{x}_{t:T})\,.$$

In other words, knowledge of the past observations is redundant for the distribution of the hidden Markov chain when we condition on its previous value. Therefore, when $\mathcal{Y}$ is finite, we can write that

$$p(y_T|y_{T-1},\mathbf{x}_{0:T}) \propto p_{y_{T-1}y_T} f(x_T|y_T) \equiv p_T^\star(y_T|y_{T-1},\mathbf{x}_{0:T})\,,$$

meaning that we define $p_T^\star(y_T|y_{T-1},\mathbf{x}_{0:T})$ as the unnormalized version of the density $p(y_T|y_{T-1},\mathbf{x}_{0:T})$ and then we can process backward for the definition of the previous conditionals, so that ($1 < t < T$)

$$\begin{aligned}
p(y_t|y_{t-1},\mathbf{x}_{0:T}) &= \sum_{i=1}^{\kappa} p(y_t, y_{t+1}=i|y_{t-1},\mathbf{x}_{t:T}) \\
&\propto \sum_{i=1}^{\kappa} p(y_t, y_{t+1}=i,\mathbf{x}_{t:T}|y_{t-1}) \\
&= \sum_{i=1}^{\kappa} p(y_t|y_{t-1}) f(x_t|y_t) p(y_{t+1}=i,\mathbf{x}_{(t+1):T}|y_t) \\
&\propto p_{y_{t-1}y_t} f(x_t|y_t) \sum_{i=1}^{\kappa} p(y_{t+1}=i|y_t,\mathbf{x}_{(t+1):T}) \\
&\propto p_{y_{t-1}y_t} f(x_t|y_t) \sum_{i=1}^{\kappa} p_{t+1}^\star(i|y_t,\mathbf{x}_{1:T}) \equiv p_t^\star(y_t|y_{t-1},\mathbf{x}_{1:T})\,.
\end{aligned}$$

At last, the conditional distribution of the starting hidden value is

$$p(y_0|\mathbf{x}_{0:T}) \propto \varrho_{y_0}\, f(x_0|y_0) \sum_{i=1}^{\kappa} p_1^\star(i|y_0,\mathbf{x}_{0:t}) \equiv p_0^\star(y_0|\mathbf{x}_{0:T})\,,$$

where $(\varrho_k)_k$ is the stationary distribution associated with the Markov transition matrix $\mathbb{P}$.

[18] To lighten notation, we will not use the parameters appearing in the various distributions of the HMM, although they are obviously of central interest.

[19] This recurrence relation has been known for quite a while in the signal processing literature and is also used in the corresponding EM algorithm; see Cappé et al. (2004) for details.

While this construction amounts to a straightforward conditioning argument, the use of the unnormalized functions $p^\star_{t+1}(y_{t+1} = i|y_t, x_{(1:T)})$ is crucial for deriving the joint conditional distribution of $y_{1:T}$ since resorting to the normalized conditionals instead would result in a useless identity.

Ⓒ Notice that, as stated above, the derivation of the $p^\star_t$'s indeed has a cost of $\mathsf{O}(T \times \kappa^2)$ since, for each t and each of the κ values of y_t, a sum of κ terms has to be computed. So, in terms of raw computational time, computing the observed likelihood does not take less time than simulating the sequence $(y_t)_{t\in\mathbb{N}}$ in the Gibbs sampler. However, the gain in using this forward–backward formula may result in subtler effects on a resulting Metropolis–Hastings algorithm, such as a better mixing of the chain of the parameters, given that we are simulating the whole vector at once.

Once we have all the conditioning functions (or *backward* equations), it is possible to simulate sequentially the hidden sequence $\mathbf{y}_{0:T}$ given $\mathbf{x}_{0:T}$ by generating first y_0 from $p(y_0|\mathbf{x}_{0:T})$, second y_1 from $p(y_1|y_0, \mathbf{x}_{0:T})$ and so on. However, there is (much) more to be done. Indeed, when considering the joint conditional distribution of $\mathbf{y}_{0:T}$ given $\mathbf{x}_{0:T}$, we have

$$\begin{aligned} p(\mathbf{y}_{0:T}|\mathbf{x}_{0:T}) &= p(y_0|\mathbf{x}_{0:T}) \prod_{t=1}^{T} p(y_t|y_{t-1}, \mathbf{x}_{0:T}) \\ &= \frac{\pi(y_1) f(x_0|y_0)}{\sum_{i=1}^{\kappa} p^\star_0(i|\mathbf{x}_{0:T})} \prod_{t=1}^{T} \frac{p_{y_{t-1}y_t} f(x_t|y_t) \sum_{i=1}^{\kappa} p^\star_{t+1}(i|y_t, \mathbf{x}_{1:T})}{\sum_{i=1}^{\kappa} p^\star_t(i|y_{t-1}, \mathbf{x}_{(1:T)})} \\ &= \pi(y_0) f(x_0|y_0) \prod_{t=1}^{T} p_{y_{t-1}y_t} f(x_t|y_t) \Big/ \sum_{i=1}^{\kappa} p^\star_1(i|\mathbf{x}_{0:T}) \end{aligned}$$

since all the other sums cancel. This joint conditional distribution immediately leads to the derivation of the observed likelihood since, by Bayes' formula,

$$f(\mathbf{x}_{0:T}) = \frac{f(\mathbf{x}_{0:T}|\mathbf{y}_{1:T})\, p(\mathbf{y}_{0:T})}{p(\mathbf{y}_{0:T}|\mathbf{x}_{0:T})} = \sum_{i=1}^{\kappa} p^\star_1(i|\mathbf{x}_{0:T})\,,$$

which is the normalizing constant of the initial conditional distribution!

A forward derivation of the likelihood can similarly be constructed. Besides the obvious construction that is symmetrical to the previous one, consider the *prediction filter*

$$\varphi_t(i) = \mathbb{P}(y_t = i|x_{1:t-1})\,,$$

with $\varphi_1(j) = \pi(j)$ (where the term *prediction* refers to the conditioning on the observations prior to time t). The *forward* equations are then given by $(t = 1, \ldots, T)$

$$\varphi_{t+1}(j) = \frac{1}{c_t} \sum_{i=1}^{\kappa} f(x_t|y_t = i)\varphi_t(i)p_{ij}\,,$$

where

$$c_t = \sum_{k=1}^{\kappa} f(x_t|y_t = k)\varphi_t(k)$$

is the normalizing constant. (This formula uses exactly the same principle as the backward equations.) Exploiting the Markov nature of the joint process $(x_t, y_t)_t$, we can then derive the log-likelihood as

$$\begin{aligned}\log p(\mathbf{x}_{1:t}) &= \sum_{r=1}^{t} \log \left[\sum_{i=1}^{\kappa} p(x_t, y_t = i|x_{1:(r-1)}) \right] \\ &= \sum_{r=1}^{t} \log \left[\sum_{i=1}^{\kappa} f(x_r|y_t = i)\varphi_r(i) \right] ,\end{aligned}$$

which also requires a $\mathrm{O}(T \times \kappa^2)$ computational time.

Ⓒ Obviously, to be able to handle the observed likelihood opens new avenues for simulation methods. For instance, the completion step (of simulating the hidden Markov chain) is no longer necessary, and Metropolis–Hastings alternatives such as random–walk proposals can be used.

⊞ Returning to **Dnadataset**, we can compute the log-likelihood (and hence the posterior up to a normalizing constant) associated with a given parameter using, for instance, the prediction filter. In that case,

$$\log p(\mathbf{x}_{1:T}) = \sum_{t=1}^{T} \log \left[\sum_{i=1}^{k} q^i_{x_t} \varphi_t(i) \right] ,$$

where $\varphi_t(j) \propto \sum_{i=1}^{2} q^i_{x_t} \varphi_t(i) p_{ij}$. This representation of the log-likelihood is used in the computation given above for the Gibbs sampler.

Furthermore, given that all parameters to be simulated are probabilities, using a normal random walk proposal in the Metropolis–Hastings algorithm is not adequate. Instead, a more appropriate proposal is based on Dirichlet distributions centered at the current value, with scale factor $\alpha > 0$; that is $(j = 1, 2)$,

$$\tilde{p}_{jj} \sim \mathscr{B}e(\alpha p_{jj}, \alpha(1 - p_{jj})) \quad \tilde{q^j} \sim \mathscr{D}(\alpha q^j_1, \ldots, \alpha q^j_4) .$$

The Metropolis–Hastings acceptance probability is then the ratio of the likelihoods over the ratio of the proposals, $f(\theta|\theta')/f(\theta'|\theta)$. Since larger values of α produce more local moves, we could test a range of values to determine the "proper" scale. However, this requires a long calibration step. Instead, the algorithm can take advantage of the different scales by picking at random for each iteration a value of α from among 1, 10, 100, 10000 or $10,0000$. (The randomness in α can then be either ignored in the computation of the proposal density f or integrated by a Rao–Blackwell argument.) For **Dnadataset**, this range of α's was wide enough since the average probability of acceptance is 0.25 and a chain $(\theta_m)_m$ started at random does converge to the same values as the Gibbs chains

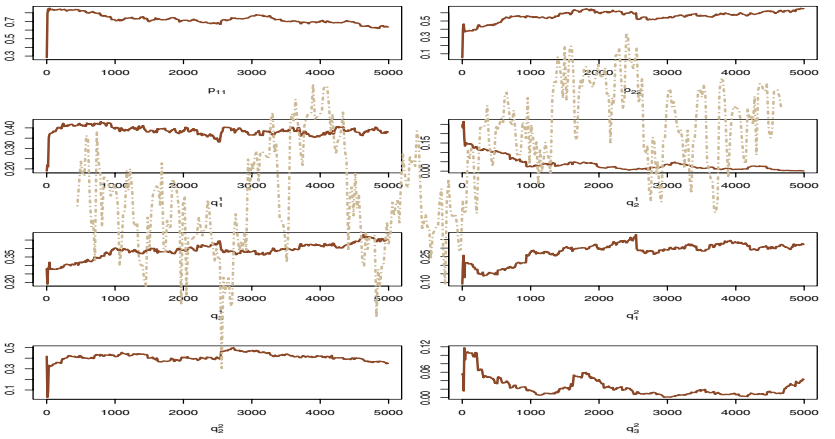


Fig. 7.10. Dataset Dnadataset: Convergence of a Metropolis–Hastings sequence for the hidden Markov model based on 5000 iterations. The overlayed curve is the sequence of log-posterior values.

simulated above, as shown by Figure 7.10, which also indicates that more iterations would be necessary to achieve complete stability. We can note in particular that the maximum log-posterior value found along the iterations of the Metropolis–Hastings algorithm is $-13,116$, which is larger than the values found in Table 7.1 for the Gibbs sampler, for parameter values of $(0.70, 0.58, 0.37, 0.011, 0.42, 0.19, 0.32, 0.42, 0.003, 0.26)$.

When the state space $\mathcal{Y}$ is finite, it may be of interest to estimate the order of the hidden Markov chain. For instance, in the case of **Dnadataset**, it may be interesting to know how many hidden coding states there are. Since the construction of the corresponding reversible jump MCMC algorithm is very much like the reversible jump algorithm for the mixture model developed in Section 6.7.2, we will not repeat its derivation. A useful reference is Cappé et al. (2004, Chapter 16) since the authors provide there an extensive discussion of the construction and properties of a reversible jump algorithm in this setting.

⊞ The model first introduced for **Dnadataset** is overly simplistic in that, at least within the coding regime, the x_t's are not independent. A more realistic modeling thus assumes that the x_t's constitute a Markov chain within each state of the hidden chain, resulting in the dependence graph of Figure 7.11. To distinguish this case from the earlier one, it is often called *Markov–switching*. This extension is much more versatile than the model of Figure 7.6, and we can hope to capture the time dependence better. However, it is far from parsimonious, as the use of different Markov transition matrices for *each* hidden state induces an explosion in the number of parameters. For instance, if there

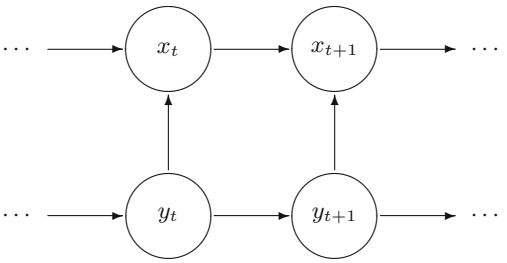


Fig. 7.11. DAG representation of the dependence structure of a Markov–switching model where $(x_t)_t$ is the observable process and $(y_t)_t$ is the hidden chain.

are two hidden states, the number of parameters is 26; if there are four hidden states, the number jumps to 60.

Exercise 7.17. Show that the counterpart of the prediction filter in the Markov-switching case is given by

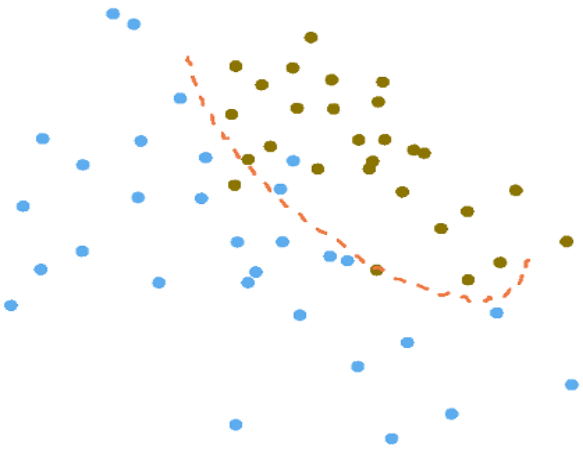
$$\log p(\mathbf{x}_{1:t}) = \sum_{r=1}^{t} \log \left[\sum_{i=1}^{\kappa} f(x_r | x_{r-1}, y_r = i) \varphi_r(i) \right] ,$$

where $\varphi_r(i) = \mathbb{P}(y_r = i | x_{1:t-1})$ is given by the recursive formula

$$\varphi_r(i) \propto \sum_{j=1}^{\kappa} p_{ji} f(x_{r-1} | x_{r-2}, y_{r-1} = j) \varphi_{r-1}(j) .$$

8

Image Analysis



"Reduce it to binary, Siobhan," she told herself.
—**Ian Rankin**, ***Resurrection Men***.—

Roadmap

This final chapter gathers two approaches to the analysis of image-related datasets, the former pertaining to the area of so-called supervised classification where some datapoints are identified as belonging to a certain group (as opposed to the unsupervised classification of mixture data in Chapter 6) and the latter pertaining to pattern detection and image correction. Classification is operated via a probabilistic version of the k-nearest-neighbor method, while pattern detection is based on Potts modeling.

Image analysis has been a very active area for both Bayesian statistics and computational methods in the past thirty years, so we feel it well deserves a chapter of its own for its specific features. This is also the only introduction to spatial statistics we will provide in this book, and we thus very briefly mention Markov random fields, which are extensions of Markov chains to the spatial domain. A complete reference on this topic is Møller (2003).

8.1 Image Analysis as a Statistical Problem

If we think of a computer image as a large collection of colored pixels, there does not seem to be any randomness involved nor any need for statistical analysis! Nonetheless, image analysis seen as a statistical analysis is a thriving field that saw the emergence of several major statistical advances, including, for instance, the Gibbs sampler. (Moreover, this field has predominantly adopted a Bayesian perspective both because this was a natural thing to do and because the analytical power of this approach was higher than with other methods.) The reason for this apparent paradox is that, while pixels usually are deterministic objects, the complexity and size of images require one to represent those pixels as the random output of a distribution governed by an object of much smaller dimension. For instance, this is the case in computer vision, where specific objects need to be extracted out of a much richer (or noisier) background.

In this spirit of extracting information from huge dimensional structure, we thus build in Section 8.2 an artificial distribution in order to relate images to predefined classes of objects, the probabilistic modeling being completely unrelated to the construction of the image itself. Similarly, in image segmentation (Section 8.3), we impose a strong spatial dimension on the prior associated with an image in order to gather homogeneous structures out of a complex or blurry image.

8.2 Computer Vision and Classification

When introduced in Chapter 6, classification was described as a way of allocating observations to different groups in order to identify homogeneous subgroups in a dataset. The technique used at that point was *unsupervised* in that the groups were at best only defined through different prior distributions, with a possible unknown number of groups. In the current setting, classification is *supervised* in the sense that it takes advantage of a richer amount of information in that some observations are already allocated to each of the groups present in the model. (This means in particular that we can use more noninformative priors on the parameters.) We first recall an ad hoc technique used in machine learning to classify datasets before presenting a more satisfactory methodology derived from this earlier technique.

8.2.1 The *k*-Nearest-Neighbor Method

The so-called *k-nearest-neighbor* procedure (abbreviated to `knn` for the corresponding R function,

```
> data(iris3)
> train=rbind(iris3[1:25,,1],iris3[1:25,,2],iris3[1:25,,3])
```

```
> test=rbind(iris3[26:50,,1],iris3[26:50,,2],iris3[26:50,,3])
> cl=factor(c(rep("s",25),rep("c",25),rep("v",25)))
> library(class)
> knn(train,test,cl,k=3,prob=TRUE)
> attributes(.Last.value)
```

as described in R online help) procedure has been used in data analysis and machine learning communities as a quick-and-dirty way to classify objects into predefined groups (Ripley, 1996). This approach requires a training dataset where both the class y and the vector $\mathbf{x}$ of characteristics (or covariates) of each observation are known. (The number G of observed classes is known and no remaining latent class can be included.) Once the training dataset has been constructed, an unsupervised observation is allocated to the class according to the highest-ranking class amidst its neighbors in the training dataset. At this stage, there is no clear difference from the generalized linear models of Chapter 4 in that a polylogit model $\mathbb{P}(y = g|\mathbf{x}) \propto \exp(\beta_g^{\mathsf{T}}\mathbf{x})$ or a polyprobit model $\mathbb{P}(y = |\mathbf{x}) \propto \Phi(\beta_g^{\mathsf{T}}\mathbf{x})$ could be used also in this setting. The difficulty however is that the dimension of the covariate vector $\mathbf{x}$ is usually too large to allow the proper derivation of the posterior distribution on the β_i's. It is often the case that the number of covariates exceeds the number of observations and this forces the statistician to build a somehow arbitrary summary measure in order to have a working prediction tool.

The training dataset $(y_i, \mathbf{x}_i)_{1\leq i\leq n}$ is used by the k-nearest-neighbor procedure to predict the value of y_{n+1} given a new vector of covariates $\mathbf{x}_{n+1}$ in a very rudimentary manner. The predicted value of y_{n+1} is simply the most frequent class found amongst the k nearest neighbors of $\mathbf{x}_{n+1}$ in the set $(\mathbf{x}_i)_{1\leq i\leq n}$. The notion of neighborhood of a point can be defined in many ways, but the usual convention is to use a distance $\rho(\cdot,\cdot)$ like the Euclidean norm. Therefore, the classical k-nearest-neighbor procedure does not involve much calibration and requires no statistical modeling at all! It can, however, be reinterpreted as a nonparametric procedure, where the value of k is chosen by minimizing a cross-validated misclassification rate. Fortunately, there exists a cleaner Bayesian reformulation, which will be detailed below.

⊞ The **vision** dataset that motivates this section is based on 1373 color pictures, described by 200 variables rather than by the whole table of 500×375 pixels. (Those variables are based on histograms of local descriptors.) There are four classes of images, depending on the main object appearing in the image as represented in Figure 8.1: Class C_1 is for motorcycles, class C_2 for bicycles, class C_3 for humans, and class C_4 for cars. This is a typical issue in computer vision problems, where one wants to build a classifier to identify a picture pertaining to a specific topic without human intervention.[1] In our processing of this dataset, we use about half of the images, namely 648 pictures, to construct the training dataset and we save the 689 remaining images to test the performance of our

[1] For instance, Web-search tools for locating images on a specific theme can be based on such principles.

classification procedures. For instance, the generic one-nearest-neighbor procedure gives a misclassification rate of roughly 30%, which amounts to 201 images misclassified.

Fig. 8.1. Dataset vision: Selected images from the four classes C_1 (motorcycles), C_2 (bicycles), C_3 (humans), and C_4 (cars).

8.2.2 A Probabilistic Version of the knn Methodology

There is a fundamental difficulty in incorporating the k-nearest-neighbor procedure within a probabilistic setup: If we try to interpret the allocation of the observation with covariate $\mathbf{x}_{n+1}$ to the highest class amongst the k nearest neighbors based on a conditional probability, there is no symmetry in the connection between the $(n+1)$st point and the n previous observations, while this is a fundamental requirement for the existence of a joint probability measure on the sample of $(n+1)$ observations (Cressie, 1993). In other words, any set of full conditional distributions based on this type of asymmetric neighborhood system is not consistent with a joint distribution.

Exercise 8.1. Draw an example with $n = 5$ points in $\mathbb{R}^2$ such that the k-nearest-neighbor relation is not symmetric.

To achieve a probabilistic interpretation of the method, we thus propose a modification of the k-nearest-neighbor structure, namely a *symmetrization* of

the neighborhood relation: If $\mathbf{x}_i$ belongs to the k-nearest-neighborhood of $\mathbf{x}_j$ and if $\mathbf{x}_j$ does not belong to the k-nearest-neighborhood of $\mathbf{x}_i$, the point $\mathbf{x}_j$ is added to the set of neighbors of $\mathbf{x}_i$. The transformed set of neighbors is then called the *symmetrized k-nearest-neighbor system.* As illustrated in Figure 8.2, the central point $\mathbf{x}_5$ has three nearest neighbors such that $\mathbf{x}_5$ is a neighbor of none of them. The symmetrized neighborhoods of those three neighbors then contain $\mathbf{x}_5$.

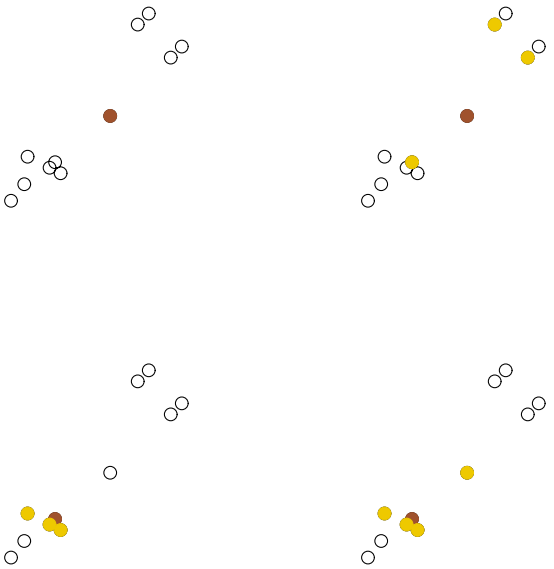


Fig. 8.2. Illustration of the symmetrization process: (*upper left*) training sample and point $\mathbf{x}_5$ (dark dot); (*upper right*) $\mathbf{x}_5$ and its three nearest neighbors; (*lower left*) $\mathbf{x}_4$ and its three nearest neighbors; (*lower right*) symmetrized three-nearest-neighborhood of $\mathbf{x}_4$ that includes $\mathbf{x}_5$.

Exercise 8.2. For a given pair (k, n) and a uniform distribution of $\mathbf{x}$ in $[0,1]^3$, design a Monte Carlo experiment that evaluates the distribution of the size of the symmetrized k-nearest-neighborhood.

Once we have defined this compatible notion of neighborhood, denoted by $i \sim_k j$ when either i is one of the k nearest neighbors of j or j is one of the k nearest neighbors of i, we can propose a complete conditional probability distribution for the class allocation. We settle for a logit-like model of the

form

$$\mathbb{P}(y_i = C_j | \mathbf{y}_{-i}, \mathbf{X}, \beta, k) = \frac{\exp\left(\beta \sum_{\ell \sim_k i} \mathbb{I}_{C_j}(y_\ell) \Big/ N_k(i)\right)}{\sum_{g=1}^{G} \exp\left(\beta \sum_{\ell \sim_k i} \mathbb{I}_{C_g}(y_\ell) \Big/ N_k(i)\right)}, \tag{8.1}$$

where $N_k(i)$ is the size of the symmetrized k-nearest-neighborhood of $\mathbf{x}_i$, $\beta > 0$, and

$$\mathbf{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)\,, \quad \mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}\,.$$

In contrast to the logit model of Chapter 4, this probability model is parameterized by the pair (β, k) and, more importantly, it is defined only by its full conditional distributions (8.1). It can be established, as discussed in Section 8.3 (Exercise 8.5), that there exists a well-defined joint distribution on $\mathbf{y} = (y_1, \ldots, y_n)$ that corresponds to this set of full conditionals, whose pdf is denoted by $f(\mathbf{y}|\mathbf{X}, \beta, k)$, but this joint distribution is missing form. The conditional distribution (8.1) is also used for the prediction of an unobserved class y_{n+1} based on the covariate vector $\mathbf{x}_{n+1}$.

Exercise 8.3. When $\mathbf{y} = (y_1, y_2)$, show that the joint pdf of $\mathbf{y}$ is given by

$$f(\mathbf{y}|\mathbf{X}, \beta, k) = f(y_1|y_2, \mathbf{X}, \beta, k) \Big/ \sum_{g=1}^{G} \frac{f(C_g|y_2, \mathbf{X}, \beta, k)}{f(y_2|C_g, \mathbf{X}, \beta, k)}\,.$$

Discuss the extension to the general case. (*Indication*: The extension is solved via the Hammersley–Clifford theorem, given in Section 8.3.1.)

The role of β in (8.1) is to grade the influence of the prevalent neighborhood class in the sense that $\beta = 0$ induces a uniform distribution on all classes, while $\beta = +\infty$ leads to a point mass at the prevalent class. The introduction of the denominator $N_k(i)$ is somehow necessary to make β dimensionless, so that the effect of a larger neighborhood is the same for a given frequency of each class. (This parameterization also reduces the dependence between k and β in the posterior distribution.) We will introduce in Section 8.3 a similar representation for the distribution of binary images called the Ising model.

The probabilistic knn model defined by (8.1) is conditional on the covariate matrix $\mathbf{X}$, but the extension of the model to future observations is based on the assumption that the frequencies $n_1/n, \ldots, n_G/n$ of the classes within the training set are representative of the marginal probabilities $p_1 = \mathbb{P}(y_i = C_1), \ldots, p_G = \mathbb{P}(y_i = C_G)$, that is, that the points $(y_i, \mathbf{x}_i)_{1 \le i \le n}$ have been generated by a multinomial $\mathcal{M}(n; p_1, \ldots, p_g)$ scheme. However, it may well be

that the training set is biased: For instance, rare or heterogeneous classes may be overrepresented in the training set to compensate for their rarity.[2] In that case, if the marginal probabilities p_g are known, a natural solution consists of reweighting the various classes according to their true frequencies. If we set the weights equal to $a_g = p_g n/n_g$ $(1 \le g \le G)$, the corrected predictive distribution of y_{n+1} is then given by

$$\mathbb{P}(y_{n+1} = C_j | \mathbf{x}_{n+1}, \mathbf{y}, \mathbf{X}, \beta, k) = \exp\left(\beta\, a_j \sum_{\ell \sim_k (n+1)} \mathbb{I}_{C_j}(y_\ell) \Big/ N_k(n+1)\right) \Big/ \sum_{g=1}^{G} \exp\left(\beta\, a_g \sum_{\ell \sim_k (n+1)} \mathbb{I}_{C_g}(y_\ell) \Big/ N_k(n+1)\right). \tag{8.2}$$

From a Bayesian perspective, given a prior distribution $\pi(\beta, k)$ with support $[0, \beta_{\max}] \times \{1, \ldots, K\}$, the marginal predictive distribution of y_{n+1} is

$$\mathbb{P}(y_{n+1} = C_j | \mathbf{x}_{n+1}, \mathbf{y}, \mathbf{X}) = \sum_{k=1}^{K} \int_0^{\beta_{\max}} \mathbb{P}(y_{n+1} = C_j | \mathbf{x}_{n+1}, \mathbf{y}, \mathbf{X}, \beta, k) \pi(\beta, k | \mathbf{y}, \mathbf{X})\, \mathrm{d}\beta\,,$$

where $\pi(\beta, k | \mathbf{y}, \mathbf{X})$ is the posterior distribution given the training dataset $(\mathbf{y}, \mathbf{X})$. There is, however, a major difficulty with this formal solution, namely that, as mentioned above, it is impossible to compute $f(y|x, \beta, k)$.[3]

An approximation often found in image analysis and other random field modeling is to use instead a *pseudo-likelihood* made of the product of the full conditionals

$$\widehat{f}(\mathbf{y} | \mathbf{X}, \beta, k) = \prod_{g=1}^{G} \prod_{y_i = C_g} \mathbb{P}(y_i = C_g | \mathbf{y}_{-i}, \mathbf{X}, \beta, k)\,.$$

This is a double approximation in that (a) $\widehat{f}(\mathbf{y}|\mathbf{X}, \beta, k)$ does not integrate to one and the normalization constant is both missing and unknown, and (b) the joint distribution corresponding to $\widehat{f}(\mathbf{y}|\mathbf{X}, \beta, k)$ does not provide the $\mathbb{P}(y_i = \cdot|\mathbf{y}_{-i}, \mathbf{X}, \beta, k)$'s as full conditionals. However, the first point is irrelevant in terms of MCMC implementation and we still use this approximation to derive both the corresponding posterior $\widehat{\pi}(\beta, k | \mathbf{y}, \mathbf{X}) \propto \widehat{f}(\mathbf{y}|\mathbf{X}, \beta, k)\pi(\beta, k)$ and the pseudo-predictive distribution

[2] This biased strategy is quite common in survey sampling—rare subpopulations are often oversampled to ensure enough representativity.

[3] More exactly, the normalizing constant of $f(y|x, \beta, k)$ cannot be constructed, although it depends on both β and k and is thus relevant for the posterior distribution.

$$\widehat{\mathbb{P}}(y_{n+1} = C_j | \mathbf{x}_{n+1}, \mathbf{y}, \mathbf{X}) = \sum_{k=1}^{K} \int_0^{\beta_{\max}} \mathbb{P}(y_{n+1} = C_j | \mathbf{x}_{n+1}, \mathbf{y}, \mathbf{X}, \beta, k) \widehat{\pi}(\beta, k | \mathbf{y}, \mathbf{X}) \, \mathrm{d}\beta \,. \tag{8.3}$$

8.2.3 MCMC Implementation

The explicit computation of (8.3) is obviously intractable and an MCMC approximation is required, based on a Markov chain $(\beta^{(1)}, k^{(1)}), \ldots, (\beta^{(M)}, k^{(M)})$ with stationary distribution $\widehat{\pi}(\beta, k | y, x)$ and whose simulation is now detailed.

Ⓒ First, we could try to use a Gibbs sampling scheme based on the fact that the posterior conditional distributions of both β and k are available; e.g., $\widehat{\pi}(\beta | k, \mathbf{y}, \mathbf{X}) \propto \widehat{f}(\mathbf{y} | \mathbf{X}, \beta, k) \pi(\beta, k)$. Since the conditional distribution of β is nonstandard, we need to resort to a hybrid scheme in which the exact simulation from $\widehat{\pi}(\beta | k, \mathbf{y}, \mathbf{X})$ is replaced with a Metropolis–Hastings step. Furthermore, using the full conditional distribution of k can induce very intensive computations. Indeed, for each new value $\beta^{(t)}$, we have to calculate the pseudo-likelihood $\widehat{f}(\mathbf{y} | \mathbf{X}, \beta^{(t)}, i) \pi(\beta^{(t)}, i)$ for $i = 1, \ldots, K$. If both n and K are large, this is quite time-consuming. Therefore, we propose to use instead a random walk Metropolis–Hastings algorithm.

Both β and k are updated using random walk proposals. Since $\beta \in [0, \beta_{\max}]$, we propose a logistic transform on β, the reparameterization

$$\beta^{(t)} = \beta_{\max} \exp(\theta^{(t)}) / (\exp(\theta^{(t)}) + 1) \,,$$

in order to be able to simulate a normal random walk on the θ's, $\tilde{\theta} \sim \mathcal{N}(\theta^{(t)}, \tau^2)$, without support restrictions. For the update of k, we use a uniform proposal on the r neighbors of $k^{(t)}$, namely on $\{k^{(t)} - r, \ldots, k^{(t)} - 1, k^{(t)} + 1, \ldots k^{(t)} + r\} \cap \{1, \ldots, K\}$. The proposal distribution $Q_r(k, \cdot)$ with pdf $q_r(k, k')$ thus depends on a scale parameter $r \in \{2, \ldots, \lfloor (K-1)/2 \rfloor\}$ that needs to be calibrated. The corresponding MCMC algorithm for estimating the parameters of the knn model is therefore as follows:

Algorithm 8.1. knn Metropolis–Hastings Sampler

Initialization: Generate $\theta^{(0)} \sim \mathcal{N}\left(0, \tau^2\right)$ and $k^{(0)} \sim \mathcal{U}_{\{1, \ldots, K\}}$.

Iteration t $(t \geq 1)$:

1. Generate $\tilde{\theta} \sim \mathcal{N}\left(\theta^{(t-1)}, \tau^2\right)$ and $\tilde{k} \sim Q_r(k^{(t-1)}, \cdot)$.
2. Compute $\tilde{\beta} = \beta_{\max} \exp(\tilde{\theta}) / (\exp(\tilde{\theta}) + 1)$ and

$$\rho = \frac{\widehat{\pi}(\tilde{\beta}, \tilde{k} | \mathbf{y}, \mathbf{X})}{\widehat{\pi}(\beta^{(t-1)}, k^{(t-1)} | \mathbf{y}, \mathbf{X})} \frac{e^{\tilde{\theta}}}{e^{\theta^{(t-1)}}} \frac{(1 + e^{\theta^{(t-1)}})^2 q_r(\tilde{k}, k^{(t-1)})}{(1 + e^{\tilde{\theta}})^2 \, q_r(k^{(t-1)}, \tilde{k})} \,.$$

3. Take

$$\left(\beta^{(t)}, \theta^{(t)}, k^{(t)}\right) = \begin{cases} \left(\tilde{\beta}, \tilde{\theta}, \tilde{k}\right) & \text{with probability } \rho \vee 1 \,, \\ \left(\beta^{(t-1)}, \theta^{(t-1)}, k^{(t-1)}\right) & \text{otherwise.} \end{cases}$$

Note that the ratio $q_r(\tilde{k}, k^{(t-1)})/q_r(k^{(t-1)}, \tilde{k})$ does not simplify to 1 in every occurrence because of boundary effects on the normalizing constant of the uniform distribution.

Using this algorithm, we can thus derive the most likely class associated with a covariate vector $\mathbf{x}_{n+1}$ from the approximated probabilities,

$$\frac{1}{M}\sum_{i=1}^{M}\widehat{\mathbb{P}}\left(y_{n+1}=l|x_{n+1},y,x,(\beta^{(i)},k^{(i)})\right).$$

This is an approximation to the MAP estimate of y_{n+1}, but, as usual, we can also ascertain the uncertainty associated with this estimation.

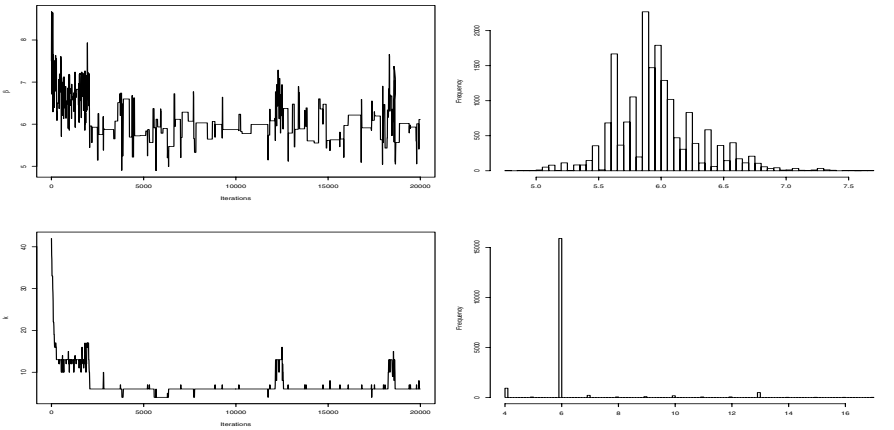


Fig. 8.3. Dataset vision: k sequences for the knn Metropolis–Hastings β based on 20,000 iterations.

⊞ For **vision**, Figure 8.3 illustrates the output from a knn Metropolis–Hastings run when $\beta_{\max} = 15$, $K = 83$, $\tau^2 = 0.05$, and $r = 1$. Note that sequences of both β's and k's are poorly mixing, but this is a result of the size of the dataset. For a smaller sample size, mixing of the Markov chain is much better, as illustrated below. (One may also blame the choice of both r and τ, which may not be appropriate for this dataset. For instance, the concentration of the $k^{(i)}$ at the value $k = 5$ may be attributed to either a very peaked posterior distribution or to a lack of energy in the proposal.) On the testing dataset, the resulting misclassification rate is about 22% (which amounts to 154 misclassified points).

We now relate the k-nearest-neighbor methodology to a classification model introduced in Chapter 4 that is not connected to image analysis. Indeed, (8.1) is a generalization of the logistic model that can be applied to any

dichotomous (or polytomous) dataset, obviously resulting in a conditional distribution that does not belong to the class of generalized linear models any longer.

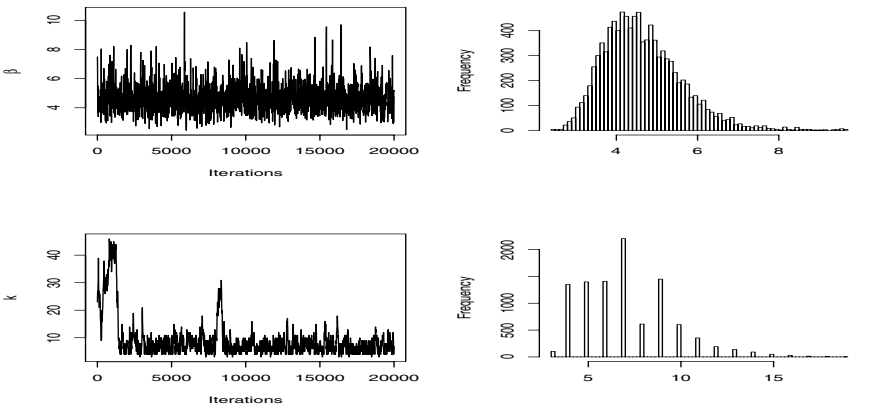


Fig. 8.4. Dataset bank: Same legend as Figure 8.3.

⊞ Recall that the dataset **bank**, introduced in Chapter 4, is made of four measurements on 100 genuine Swiss banknotes and 100 counterfeit ones. We now split this dataset at random into two equal parts: 100 banknotes are allocated to the training dataset and the remaining 100 banknotes to the testing dataset. (We do not take into account the obvious fact that the proportion of forged banknotes is not in the same proportion as the proportion of genuine banknotes in Switzerland (!) because, by construction, both the training and the testing datasets have the same proportions.) Figure 8.4 provides the equivalent of Figure 8.3 for the bank dataset, using $\beta_{\max} = 15$, $K = 50$, $\tau^2 = 0.05$, and $r = 1$. The comparison shows that the mixing has improved substantially, with a good acceptance rate for the β's and reasonable mixing for the k's (in the sense that the sequence moves over the interval $(4, 12)$). On the test dataset, the misclassification rate is now 15%.

We can also compare this classification result with the one based on a Bayesian probit model with a flat prior distribution. For the training dataset, Figure 8.5 summarizes the output of an MCMC experiment on the Bayesian estimation of the probit coefficients, using only the training dataset. (All indicators provided by this figure are good since mixing, histogram regularity, and limited autocorrelation are satisfactory for the four coefficients.) Based on the chain thus produced, we can derive the predictive probability that a bill in the test dataset is a counterfeit. The approximation to this predictive probability is obtained in a coherent way, namely, using the average of the terms

$$\Phi\left(\beta_1^{(t)} x_{(n+1)1} + \beta_2^{(t)} x_{(n+1)2} + \beta_3^{(t)} x_{(n+1)3} + \beta_4^{(t)} x_{(n+1)4}\right) ,$$

rather than resorting to crude plug-in estimates.[4] Just as in the knn approach, we can allocate each bill to the corresponding most likely group. This classification results in a 13% misclassification rate, so that the probit model does slightly better here than the knn classifier. The advantage to the probit model however has to be downweighted by the comparatively higher complexity of a model that involves four parameters to operate (instead of a fixed two for the knn model).

8.3 Image Segmentation

In this section, we still consider images as statistical objects, but they are now "noisy" in the sense that the color or the grey level of a pixel is not observed exactly but with some perturbation (sometimes called *blurring* as in satellite imaging). The purpose of image segmentation is to cluster pixels into homogeneous classes without supervision or preliminary definition of those classes, based only on the spatial coherence of the structure.

This underlying structure of the "true" pixels is denoted by $\mathbf{x}$, while the observed image is denoted by $\mathbf{y}$. Both objects $\mathbf{x}$ and $\mathbf{y}$ are arrays, with each entry of $\mathbf{x}$ taking a finite number of values and each entry of $\mathbf{y}$ taking real values (for modeling convenience rather than reality constraints). We are thus interested in the posterior distribution of $\mathbf{x}$ given $\mathbf{y}$ provided by Bayes' theorem, $\pi(\mathbf{x}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})$. In this posterior distribution, the likelihood, $f(\mathbf{y}|\mathbf{x})$, describes the link between the observed image and the underlying classification; that is, it gives the distribution of the noise, while the prior $\pi(\mathbf{x})$ encodes beliefs about the (possible or desired) properties of the underlying image. Since the object to be modeled is a random array and is thus more complex than in earlier instances, we now devote some space to the modeling used on $\mathbf{x}$: The prior will be constructed in Sections 8.3.1 and 8.3.2, while inference per se is described in Section 8.3.3. Although, as in other chapters, we cannot provide the full story of Bayesian image segmentation, an excellent tutorial on Bayesian image processing based on a summer school course can be found in Hurn et al. (2003).

As indicated above, a proper motivation for image segmentation is satellite processing since images caught by satellites are often blurred, either because of inaccuracies in the instruments or transmission or because of clouds or vegetation cover between the satellite and the area of interest.

⊞ The **Menteith** dataset that motivates this section is a 100×100 pixel satellite image of the lake of Menteith, as represented in Figure 8.6. The lake of Menteith is located in

[4] We stress once more that plug-in estimates are both wrong and useless, the former because they miss the variability due to the estimate and the latter because an MCMC output can be used at about the same cost to produce a more accurate approximation to the genuine Bayes estimate. Interested readers can check in Robert (2001, Chapter 10) for a similar discussion about the shortcomings of empirical Bayes analysis.

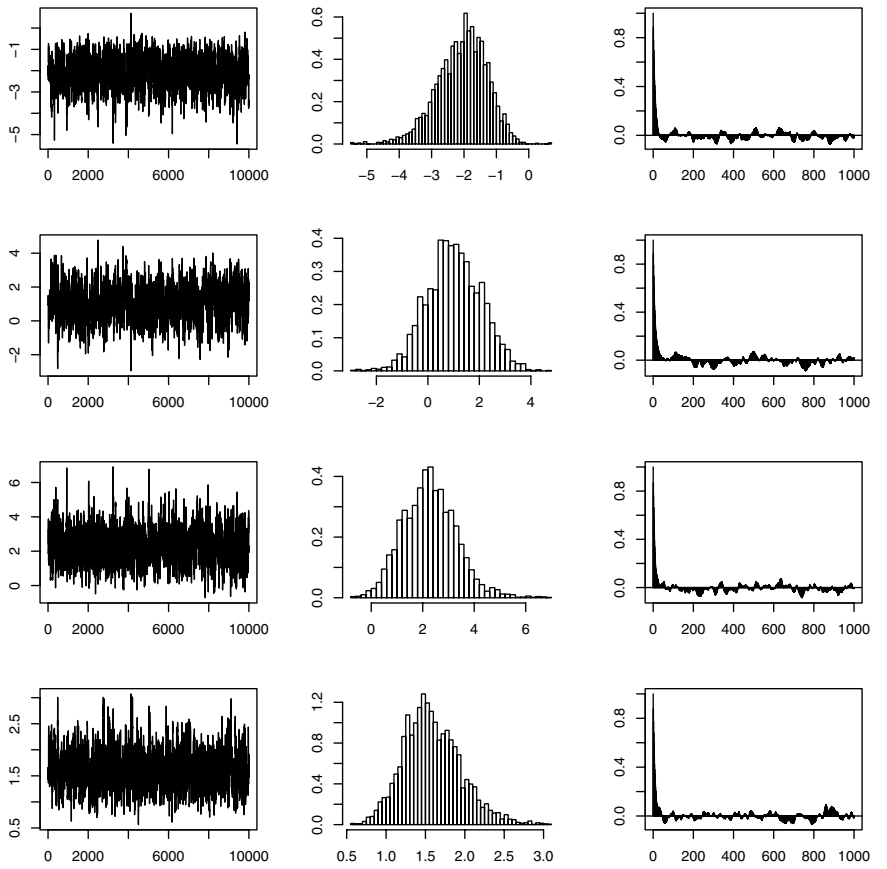


Fig. 8.5. Dataset bank: Estimation of the probit coefficients for the training dataset using Algorithm 4.2 and a flat prior.

Scotland, near Stirling, and offers the peculiarity of being called "lake" rather than the traditional Scottish "loch." As can be spotted on the image, there are several islands in this lake, one of which contains an ancient abbey. The purpose of analyzing this satellite dataset is to classify all pixels into one of six states in order to detect some homogeneous regions.

8.3.1 Markov Random Fields

In order to properly describe a structure in the original image, we need to expand the notion of Markov chain to an array (or equivalently to any lattice[5]

[5] An equivalent term for *lattice* is grid.

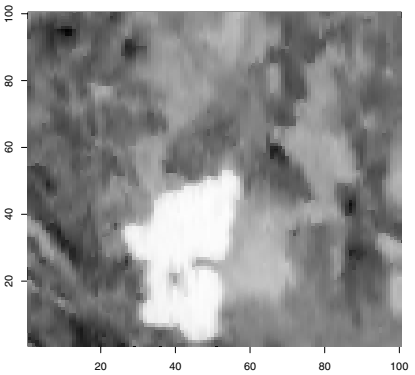


Fig. 8.6. Dataset Menteith: Satellite image of the lake of Menteith.

object). Since this is a multidimensional object, a first requirement for the generalization is to define a proper neighborhood structure.

If we take a lattice $\mathcal{I}$ of sites or pixels in an image ($\mathcal{I}$ is therefore most often an array), we denote by $i \in \mathcal{I}$ a coordinate in the lattice. The neighborhood relation is then denoted by $\sim$. For reasons already mentioned in Section 8.2, we force this relation to be symmetric: If i is a neighbor of j (written as $i \sim j$), then j is a neighbor of i. (By convention, i is not a neighbor of itself.) Figure 8.7 illustrates this notion for three types of neighborhoods on a regular grid.

A *random field* on $\mathcal{I}$ is a random structure indexed by the lattice $\mathcal{I}$, a collection of random variables $\{x_i; i \in \mathcal{I}\}$ where each x_i takes values in a finite set χ. Obviously, the interesting case is when the x_i's are dependent random variables.

If $n(i)$ is the set of neighbors of $i \in \mathcal{I}$ and if $\mathbf{x}_A = \{x_i; i \in A\}$ denotes the subset of $\mathbf{x}$ for indices in a subset $A \subset \mathcal{I}$, then $\mathbf{x}_{n(i)}$ is the set of values taken by the neighbors of i. The extension of a Markov chain to this kind of object then assumes only dependence on the neighbors.[6] If, as before, we denote by $\mathbf{x}_{-A} = \{x_i; i \notin A\}$ the coordinates that are *not* in a given subset $A \subset \mathcal{I}$, a random field is a *Markov random field* (MRF) if the conditional distribution of any pixel given the other pixels only depends on the values of the neighbors of that pixel; i.e., for $i \in \mathcal{I}$,

$$\pi(x_i|\mathbf{x}_{-i}) = \pi(x_i|\mathbf{x}_{n(i)}) \,.$$

Markov random fields have been used for quite a while in imaging, not necessarily because images obey Markovian laws but rather because these dependence structures offer highly stabilizing properties in modeling. Indeed, constructing the joint prior distribution of an image is a daunting task because

[6] This dependence obviously forces the neighborhood relation to be symmetric.

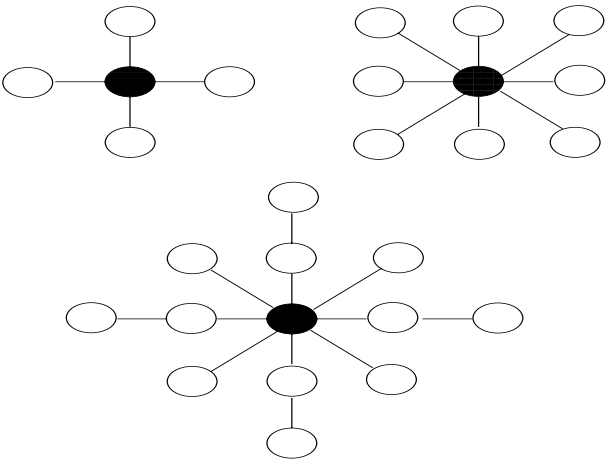


Fig. 8.7. Some common neighborhood structures in imaging: Four (*upper left*), eight (*upper right*) or twelve nearest neighbors (*lower*).

there is no immediate way of describing the global properties of an image via a probability distribution. Using the full conditional distributions breaks the problem down to a sequence of *local* problems and is therefore more manageable in the sense that we may be able to express more clearly how we think x_i behaves when the configuration of its neighbors is known.[7]

Before launching into the use of specific MRFs to describe prior assumptions on a given image (or lattice), we need to worry[8] about the very existence of MRFs! Indeed, defining a set of full conditionals does not guarantee that there is a joint distribution behind them (Exercise 8.4). In our case, this means that general forms of neighborhoods and general types of dependences on the neighbors do not usually correspond to a joint distribution on $\mathbf{x}$.

Exercise 8.4. Find two conditional distributions $f(x|y)$ and $g(y|x)$ such that there is no joint distribution corresponding to both f and g. Find a necessary condition for f and g to be compatible in that respect; i.e., to correspond to a joint distribution on (x, y).

Starting from a complete set of full conditionals on a lattice $\mathcal{I}$, if there exists a corresponding joint distribution, $\pi(\mathbf{x})$, it is sufficient to find $\pi(\mathbf{x})/\pi(\mathbf{x}^*)$

[7] It is no surprise that computational techniques such as the Gibbs sampler stemmed from this area, as the use of conditional distributions is deeply ingrained in the imaging community.

[8] For those that do not want to worry, the end of this section can be skipped, it being of a more theoretical nature and not used in the rest of the chapter.

for a given fixed value $\mathbf{x}^*$ since the normalizing constant is automatically determined. Now, if $\mathcal{I} = \{1, \ldots, n\}$, it is simple to exhibit a full conditional density within the joint density by writing the natural decomposition

$$\pi(\mathbf{x}) = \pi(x_1|\mathbf{x}_{-1})\pi(\mathbf{x}_{-1})$$

and then to introduce $\mathbf{x}^*$ by the simple divide-and-multiply trick

$$\pi(\mathbf{x}) = \frac{\pi(x_1|\mathbf{x}_{-1})}{\pi(x_1^*|\mathbf{x}_{-1})}\,\pi(x_1^*, \mathbf{x}_{-1})\,.$$

If we iterate this trick for all terms in the lattice (assuming we never divide by 0), we eventually get to the representation

$$\frac{\pi(\mathbf{x})}{\pi(\mathbf{x}^*)} = \prod_{i=0}^{n-1} \frac{\pi(x_{i+1}|x_1^*, \ldots, x_i^*, x_{i+2}, \ldots, x_n)}{\pi(x_{i+1}^*|x_1^*, \ldots, x_i^*, x_{i+2}, \ldots, x_n)}\,. \tag{8.4}$$

Hence, we can truly write the joint density as a product of ratios of its full conditionals modulo one renormalization.[9]

This result can also be used toward our purpose: If there exists a joint density such that the full conditionals never cancel, then (8.4) must hold for every representation of $\mathcal{I} = \{1, \ldots, n\}$; that is, for every ordering of the indices, and for every choice of reference value $\mathbf{x}^*$. Although we cannot provide here the reasoning behind the result, there exists a necessary and sufficient condition for the existence of an MRF. This condition relies on the notion of *clique*: Given a lattice $\mathcal{I}$ and a neighborhood relation $\sim$, a clique is a maximal subset of $\mathcal{I}$ made of sites that are all neighbors. The corresponding result (Cressie, 1993) is that an MRF associated with $\mathcal{I}$ and $\sim$ is the neighborhood relation necessarily of the form

$$\pi(\mathbf{x}) \propto \exp\left(-\sum_{C \in \mathcal{C}} \Phi_C(\mathbf{x}_C)\right)\,, \tag{8.5}$$

where $\mathcal{C}$ is the collection of all cliques. This result amounts to saying that the joint distribution must separate in terms of its system of cliques.

Exercise 8.5. Using the Hammersley–Clifford theorem, show that the full conditional distributions given by (8.1) are compatible with a joint distribution.

Exercise 8.6. If $\pi(x_1, \ldots, x_n)$ is a density such that its full conditionals never cancel on its support, characterize the support of π in terms of the supports of the marginal distributions.

[9] This representation is by no means limited to MRFs: it holds for every joint distribution such that the full conditionals never cancel. It is called the *Hammersley–Clifford theorem*, and a two-dimensional version of it was introduced in Exercise 3.21.

Exercise 8.7. Describe the collection of cliques $\mathcal{C}$ for an eight-neighbor neighborhood structure such as in Figure 8.7 on a regular $n \times m$ array. Compute the number of cliques.

8.3.2 Ising and Potts Models

We now embark on the description of two specific MRFs that are appropriate for image analysis, namely the *Ising model* used for binary images and its extension, the *Potts model*, used for images with more than two colors.

If pixels of the underlying (true) image $\mathbf{x}$ can only take two colors (black and white, say), $\mathbf{x}$ is then binary, while $\mathbf{y}$ is a grey-level image. We typically refer to each pixel x_i as being *foreground* if $x_i = 1$ (black) and *background* if $x_i = 0$ (white). The conditional distribution of a pixel is then Bernoulli, and its probability parameter can be constructed as a function of the number of black neighboring pixels, using for instance a logit link as in (8.1), namely, $(j = 0, 1)$

$$\pi(x_i = j|\mathbf{x}_{-i}) \propto \exp(\beta n_{i,j}), \qquad \beta > 0,$$

where $n_{i,j} = \sum_{\ell \in n(i)} \mathbb{I}_{x_\ell = j}$ is the number of neighbors of x_i with color j. The *Ising model* is then defined via these full conditionals

$$\pi(x_i = 1|\mathbf{x}_{-i}) = \frac{\exp(\beta n_{i,1})}{\exp(\beta n_{i,0}) + \exp(\beta n_{i,1})},$$

and the joint distribution therefore satisfies

$$\pi(\mathbf{x}) \propto \exp\left(\beta \sum_{j \sim i} \mathbb{I}_{x_j = x_i}\right), \tag{8.6}$$

where the summation is taken over all pairs (i, j) of neighbors.

Exercise 8.8. Use the Hammersley–Clifford theorem to establish that (8.6) is the joint distribution associated with the above conditionals. Deduce that the Ising model is an MRF.

When simulating a posterior distribution β in Section 8.3.3, we will be faced with a major problem with this model, namely that the normalizing constant of (8.6), $Z(\beta)$, is intractable except for very small lattices $\mathcal{I}$, while depending on β. At this stage, however, we consider β to be fixed and focus on the simulation of $\mathbf{x}$ in preparation for the inference on both β and $\mathbf{x}$ given $\mathbf{y}$.

Exercise 8.9. Draw the function $Z(\beta)$ for a 3×5 array. Determine the computational cost of the derivation of the normalizing constant $Z(\beta)$ of (8.6) for an $m \times n$ array.

Due to the convoluted correlation structure of the Ising model, direct simulation of $\mathbf{x}$ (and to an even greater extent simulation conditional on $\mathbf{y}$) is not possible. Faced with this difficulty, very early the image community developed computational tools that eventually led to the proposal of the Gibbs sampler (Section 3.4.1).[10] By construction, full conditional distributions of the Ising model are naturally available, and the local structure of Markov random fields provides an immediate single-site update for the Gibbs sampler, as shown in Algorithm 8.2.

ALGORITHM 8.2. ISING GIBBS SAMPLER

Initialization: For $i \in \mathcal{I}$, generate independently

$$x_i^{(0)} \sim \mathscr{B}(1/2)\,.$$

Iteration t $(t \geq 1)$:

1. Generate $\mathbf{u} = (u_i)_{i \in \mathcal{I}}$, a random ordering of the elements of $\mathcal{I}$.
2. For $1 \leq \ell \leq |\mathcal{I}|$, update $n_{u_\ell,0}^{(t)}$ and $n_{u_\ell,1}^{(t)}$, and generate

$$x_{u_\ell}^{(t)} \sim \mathscr{B}\left\{ \frac{\exp(\beta n_{u_\ell,1}^{(t)})}{\exp(\beta n_{u_\ell,0}^{(t)}) + \exp(\beta n_{u_\ell,1}^{(t)})} \right\}.$$

© In this implementation, the order of the updates of the pixels of $\mathcal{I}$ is random in order to overcome possible bottlenecks in the exploration of the distribution, although this is not a necessary condition for the algorithm to converge. In fact, when considering two pixels x_1 and x_2 that are m pixels away, the influence of a change in x_1 is not felt in x_2 before at least m iterations of the Gibbs sampler. Of course, if m is large, the dependence between x_1 and x_2 is quite moderate, but this slow propagation of changes is indicative of slow mixing in the Markov chain. For instance, to see a change of color of a relatively large homogeneous region is an event of very low probability, even though the distribution of the colors is exchangeable. If β is large enough, the Gibbs sampler will face enormous difficulties to simply change the value of a single pixel.

[10] The very name "Gibbs sampling" was proposed in reference to Gibbs random fields, related to the physicist Willard Gibbs. Both of the major MCMC algorithms are thus named after physicists and were originally developed for problems that were beyond the boundary of statistical inference.

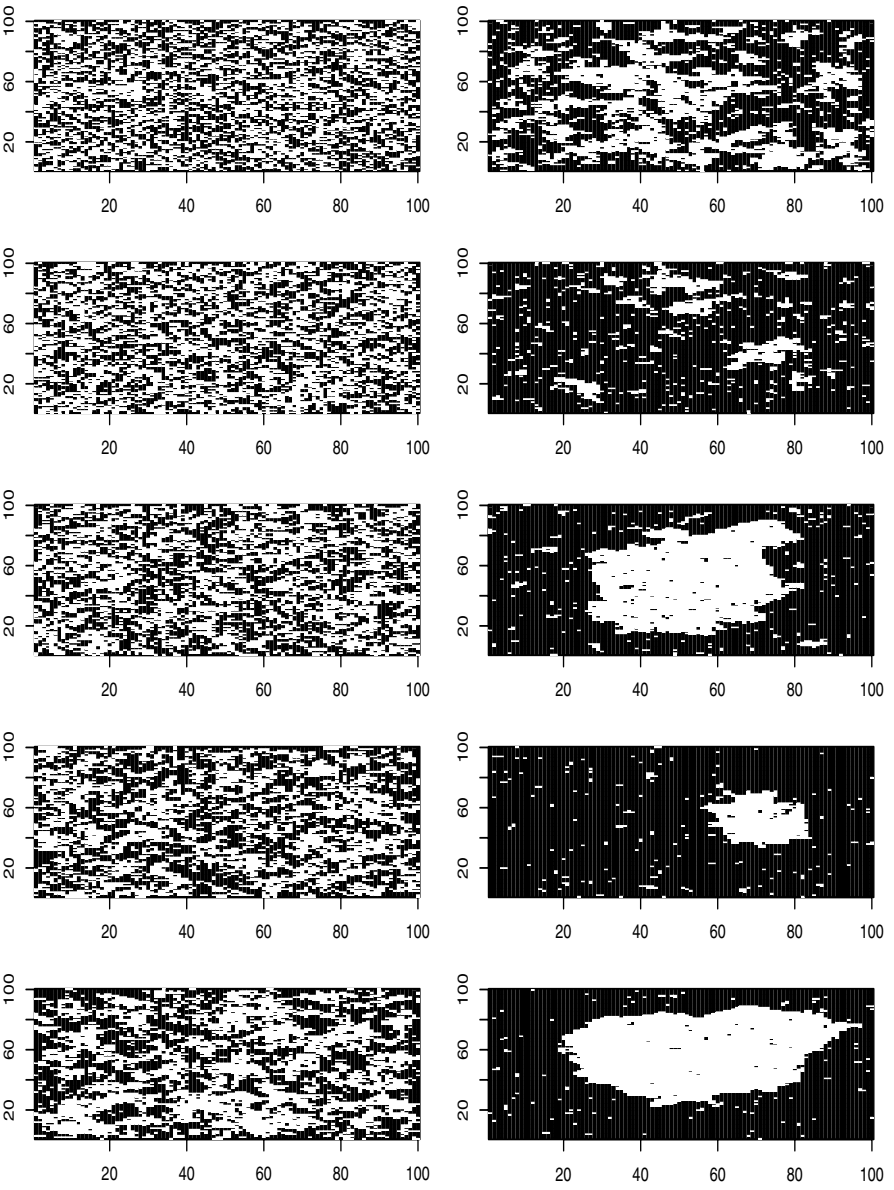


Fig. 8.8. Simulations from the Ising model with a four-neighbor neighborhood structure on a 100×100 array after $1,000$ iterations of the Gibbs sampler: β varies in steps of 0.1 from 0.3 to 1.2 (*first column, then second column*).

Figure 8.8 presents the output of simulations from Algorithm 8.2 for different values of β. Although we cannot discuss here convergence assessment for the Gibbs sampler (see Robert and Casella, 2004, Chapter 12), the images thus produced are representative of the Ising distributions: the larger β, the more homogeneous the image (and also the slower the Gibbs sampler).[11] When looking at the result associated with the larger values of β, we can start to see the motivations for using such representations to model images like the Menteith dataset.

Along with the slow dynamic induced by the single-site updating, we can point out another inefficiency of this algorithm, namely that many updates will not modify the current value of $\mathbf{x}$ simply because the new value of x_l will equal its previous value! It is, however, straightforward to modify the algorithm in order to only propose changes of values. The update of each pixel l is then a Metropolis–Hastings step with acceptance probability $\rho = \exp(\beta n_{l,1-x_l})/\exp(\beta n_{l,x_l}) \wedge 1$.

Exercise 8.10. For an $n \times m$ array $\mathcal{I}$, if the neighborhood relation is based on the four nearest neighbors, show that the $x_{2i,2j}$'s are independent conditional on the $x_{2i-1,2j-1}$'s ($1 \leq i \leq n$, $1 \leq j \leq m$). Deduce that the update of the whole image can be done in two steps by simulating the pixels with even indices and then the pixels with odd indices. (This modification of Algorithm 8.2 is a version of *the Swendsen–Wang* algorithm.)

The generalization to the case when the image has more than two colors is immediate. If there are G colors and if $n_{i,g}$ denotes the number of neighbors of $i \in \mathcal{I}$ with color g ($1 \leq g \leq G$) (that is, $n_{i,g} = \sum_{j \sim i} \mathbb{I}_{x_j = g}$), the full conditional distribution of x_i is chosen to satisfy $\pi(x_i = g | \mathbf{x}_{-i}) \propto \exp(\beta n_{i,g})$. This choice corresponds to the *Potts model*, whose joint density is given by

$$\pi(\mathbf{x}) \propto \exp\left(\beta \sum_{j \sim i} \mathbb{I}_{x_j = x_i}\right) . \tag{8.7}$$

This model suffers from the same drawback as the Ising model in that the normalizing constant is not available in closed form.

Exercise 8.11. Determine the computational cost of the derivation of the normalizing constant of the distribution (8.7) for an $m \times n$ array and G different colors.

[11] In fact, there exists a critical value of β, $\beta_c = 2.269185$, such that, when $\beta > \beta_c$, the Markov chain converges to one of two different stationary distributions, depending on the starting point. In other words, the chain is no longer irreducible. In particle physics, this phenomenon is called *phase transition*.

Exercise 8.12. Use the Hammersley–Clifford theorem to establish that (8.7) is the joint distribution associated with the conditionals above. Deduce that the Potts model is an MRF.

When simulating $\mathbf{x}$ from a Potts model, the single-site Gibbs sampler may be quite slow and more efficient alternatives are available, including the Swendsen–Wang algorithm (Exercise 8.10). For instance, Algorithm 8.3 below is a Metropolis–Hastings algorithm that forces moves on the current values. (While the proposal is not a random walk, using a uniform distribution on the $G-1$ other values leads to an acceptance probability that is equal to the ratio of the target densities.)

Algorithm 8.3. Potts Metropolis–Hastings Sampler

Initialization: For $i \in \mathcal{I}$, generate independently

$$x_i^{(0)} \sim \mathscr{U}(\{1,\ldots,G\})\,.$$

Iteration t ($t \geq 1$):

1. Generate $\mathbf{u} = (u_i)_{i\in\mathcal{I}}$ a random ordering of the elements of $\mathcal{I}$.
2. For $1 \leq \ell \leq |\mathcal{I}|$,
   generate

$$\tilde{x}_{u_\ell}^{(t)} \sim \mathscr{U}(\{1, x_{u_\ell}^{(t-1)}-1, x_{u_\ell}^{(t-1)}+1,\ldots,G\})\,,$$

   compute the $n_{u_l,g}^{(t)}$ and

$$\rho_l = \left\{\exp(\beta n_{u_\ell,\tilde{x}}^{(t)})/\exp(\beta n_{u_\ell,x_{u_\ell}}^{(t)})\right\} \wedge 1\,,$$

   and set $x_{u_\ell}^{(t)}$ equal to $\tilde{x}_{u_\ell}$ with probability ρ_l.

Exercise 8.13. Derive an alternative to Algorithm 8.3 where the probabilities in the multinomial proposal are proportional to the numbers of neighbors $n_{u_\ell,g}$ and compare its performance with that of Algorithm 8.3.

Figure 8.9 illustrates the result of a simulation using Algorithm 8.3 in a situation where there are $G = 4$ colors. Notice the reinforced influence of large β's: Not only is the homogeneity higher, but there also is a larger differentiation in the colors.[12] We stress that, while β in Figure 8.9 ranges

[12] Similar to the Ising model of footnote 11, there also are a phase transition phenomenon and a critical value for β in this case.

over the same values as in Figure 8.8, the β's are not comparable since the larger number of classes in the Potts model induces a smaller value of the $n_{i,g}$'s for the neighborhood structure.

8.3.3 Posterior Inference

The model now having been fully introduced, we turn to the central issue, namely how to draw inference on the "true" image, $\mathbf{x}$, given an observed noisy image, $\mathbf{y}$. The prior on $\mathbf{x}$ is a Potts model with G categories,

$$\pi(\mathbf{x}|\beta) = \frac{1}{Z(\beta)} \exp\left(\beta \sum_{i\in\mathcal{I}} \sum_{j\sim i} \mathbb{I}_{x_j=x_i}\right) ,$$

where $Z(\beta)$ is the normalizing constant. Given $\mathbf{x}$, we assume that the observations in $\mathbf{y}$ are independent normal random variables,

$$f(\mathbf{y}|\mathbf{x}, \sigma^2, \mu_1, \ldots, \mu_G) = \prod_{i\in\mathcal{I}} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{x_i})^2\right\} .$$

This modeling is slightly approximate in that the y_i's are integer grey levels that vary between 0 and 255, but it is easier to handle than a parameterized distribution on $\{0, \ldots, 255\}$.

This setting is clearly reminiscent[13] of the mixture and hidden Markov models of Chapters 6 and 7 in that a Markovian structure, the Markov random field, is only observed through random variables indexed by the states.

In this problem, the parameters $\beta, \sigma^2, \mu_1, \ldots, \mu_G$ are usually considered *nuisance* parameters, which justifies the use of uniform priors such as

$$\begin{aligned} \beta &\sim \mathcal{U}([0, 2]) , \\ \boldsymbol{\mu} = (\mu_1, \ldots, \mu_G) &\sim \mathcal{U}(\{\boldsymbol{\mu};\, 0 \le \mu_1 \le \ldots \le \mu_G \le 255\}) , \\ \pi(\sigma^2) &\propto \sigma^{-2}\mathbb{I}_{]0,\infty[}(\sigma^2) , \end{aligned}$$

the last prior corresponding to a uniform prior on $\log\sigma$.

The upper bound on β is chosen for a very precise reason: When $\beta \ge 2$, the Potts model is almost surely concentrated on single-color images. It is thus pointless to consider larger values of β. The ordering in the μ_g's is not necessary, strictly speaking, but it avoids the label switching phenomenon

[13] Besides image segmentation, another typical illustration of such structures is *character recognition* where a machine scans handwritten documents, e.g., envelopes, and must infer a sequence of symbols (i.e., numbers or letters) from digitized pictures. Hastie et al. (2001) provides an illustration of this problem.

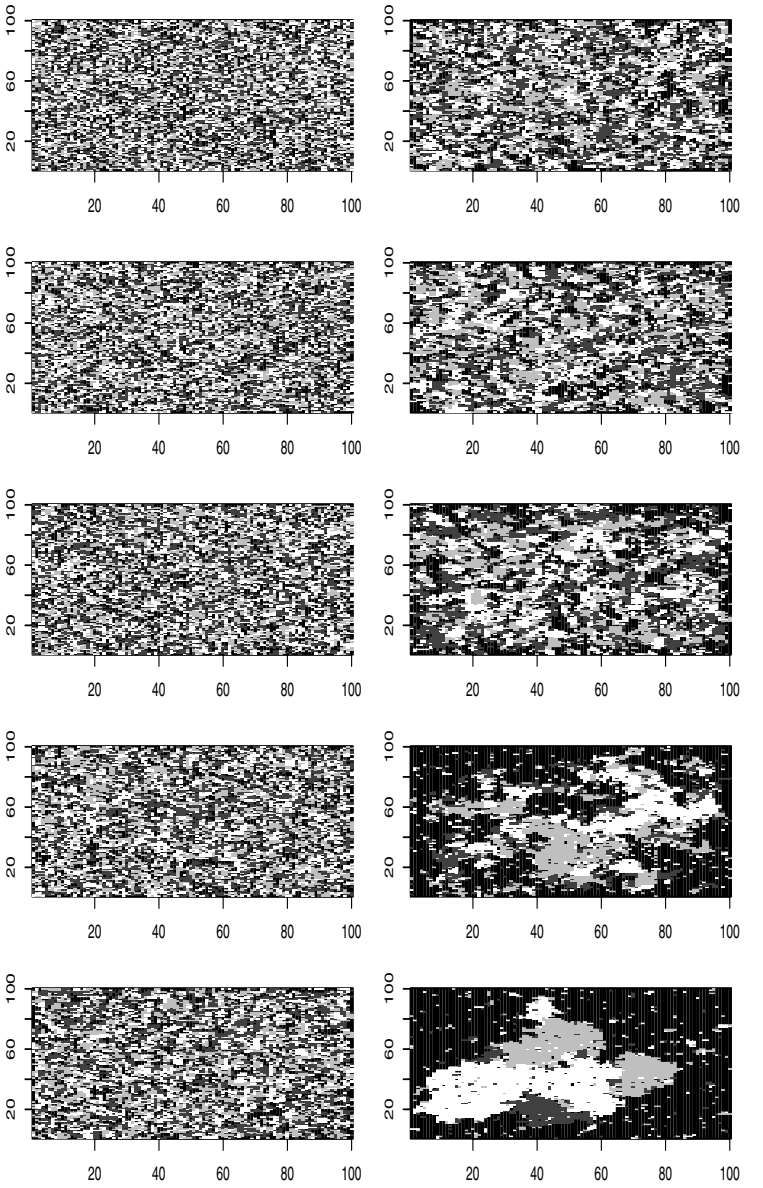


Fig. 8.9. Simulations from the Potts model with four grey levels and a four-neighbor neighborhood structure based on 1000 iterations of the Metropolis–Hastings sampler. The parameter β varies in steps of 0.1 from 0.3 to 1.2 (*first column, then second column*).

discussed in Section 6.4. (The alternative is to use the same uniform prior on all μ_g's and then reorder them once the MCMC simulation is done. While this may avoid slow convergence behaviors in some cases, this strategy also implies more involved bookkeeping and higher storage requirements. In the case of large images, it simply cannot be considered.)

The corresponding posterior distribution is thus

$$\pi(\mathbf{x}, \beta, \sigma^2, \boldsymbol{\mu}|\mathbf{y}) \propto \pi(\beta, \sigma^2, \boldsymbol{\mu}) \times \frac{1}{Z(\beta)} \exp\left(\beta \sum_{i\in\mathcal{I}} \sum_{j\sim i} \mathbb{I}_{x_j=x_i}\right) \times \prod_{i\in\mathcal{I}} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{\frac{-1}{2\sigma^2}(y_i - \mu_{x_i})^2\right\},$$

where $Z(\beta)$ denotes the normalizing constant of the Potts model.

We can therefore construct the various full conditionals of this joint distribution with a view to the derivation of a hybrid Gibbs sampler for this model. First, the full conditional distribution of x_i ($i \in \mathcal{I}$) is ($1 \le g \le G$)

$$\mathbb{P}(x_i = g|\mathbf{y}, \beta, \sigma^2, \boldsymbol{\mu}) \propto \exp\left\{\beta \sum_{j\sim i} \mathbb{I}_{x_j=g} - \frac{1}{2\sigma^2}(y_i - \mu_g)^2\right\},$$

which can be simulated directly, even though this is no longer a Potts model. As in the mixture and hidden Markov cases, once $\mathbf{x}$ is known, the groups associated with each category g separate and therefore the μ_g's can be simulated independently conditional on $\mathbf{x}$, $\mathbf{y}$, and σ^2. More precisely, if we denote by

$$n_g = \sum_{i\in\mathcal{I}} \mathbb{I}_{x_i=g} \quad \text{and} \quad s_g = \sum_{i\in\mathcal{I}} \mathbb{I}_{x_i=g} y_i$$

the number of observations and the sum of the observations allocated to group g, respectively, the full conditional distribution of μ_g is a truncated normal distribution on $[\mu_{g-1}, \mu_{g+1}]$ (setting $\mu_0 = 0$ and $\mu_{G+1} = 255$) with mean s_g/n_g and variance σ^2/n_g. (Obviously, if no observation is allocated to this group, the conditional distribution turns into a uniform distribution on $[\mu_{g-1}, \mu_{g+1}]$.) The full conditional distribution of σ^2 is an inverse gamma distribution with parameters $|\mathcal{I}|^2/2$ and $\sum_{i\in\mathcal{I}}(y_i - \mu_{x_i})^2/2$. Finally, the full conditional distribution of β is such that

$$\pi(\beta|\mathbf{y}) \propto \frac{1}{Z(\beta)} \exp\left(\beta \sum_{i\in\mathcal{I}} \sum_{j\sim i} \mathbb{I}_{x_j=x_i}\right), \tag{8.8}$$

since β does not depend on σ^2, $\boldsymbol{\mu}$, and $\mathbf{y}$, given $\mathbf{x}$.

Exercise 8.14. Show that the Swendsen–Wang improvement given in Exercise 8.10 also applies to the simulation of $\pi(\mathbf{x}|\mathbf{y}, \beta, \sigma^2, \boldsymbol{\mu})$.

Ⓒ The full conditional distribution of β is thus the only nonstandard distribution and, to run a Gibbs sampler on this model, we must resort to a hybrid scheme in that we must replace the simulation from (8.8) with some Metropolis–Hastings algorithm associated with (8.8). However, this solution runs into an obvious difficulty because of the unavailability of the normalizing constant $Z(\beta)$; a Metropolis–Hastings algorithm does need to evaluate the ratio $Z(\tilde{\beta})/Z(\beta)$. We now describe a practical resolution of this difficulty, which, while being costly in computing time, can still be implemented for large images.

Formally, the normalizing constant $Z(\beta)$ is defined as

$$Z(\beta) = \sum_{\mathbf{x}} \exp\{\beta S(\mathbf{x})\} ,$$

where $S(\mathbf{x}) = \sum_{i\in\mathcal{I}} \sum_{j\sim i} \mathbb{I}_{x_j=x_i}$ and the summation is made over the $G^{|\mathcal{I}|}$ values of $\mathbf{x}$. Since

$$\frac{\mathrm{d}Z(\beta)}{\mathrm{d}\beta} = \sum_{\mathbf{x}} S(\mathbf{x}) \exp(\beta S(\mathbf{x})) ,$$

we can express this derivative as an expectation under $\pi(\mathbf{x}|\beta)$,

$$\begin{aligned}\frac{\mathrm{d}Z(\beta)}{\mathrm{d}\beta} &= Z(\beta) \sum_{\mathbf{x}} S(\mathbf{x}) \frac{\exp(\beta S(\mathbf{x}))}{Z(\beta)} \\ &= Z(\beta)\, \mathbb{E}_\beta[S(\mathbf{X})] ,\end{aligned}$$

that is,

$$\frac{\mathrm{d}\log Z(\beta)}{\mathrm{d}\beta} = \mathbb{E}_\beta[S(\mathbf{X})] .$$

Therefore, the ratio $Z(\beta_1)/Z(\beta_0)$ can be derived from an integral, since

$$\log\{Z(\beta_1)/Z(\beta_0)\} = \int_{\beta_0}^{\beta_1} \mathbb{E}_\beta[S(\mathbf{x})]\mathrm{d}\beta , \tag{8.9}$$

which is called the *path sampling identity*. (See Chen et al., 2000, for details about this technique.)

Ⓒ Although this may not look like a considerable improvement, since we now have to compute an expectation in $\mathbf{x}$ plus an integral over β, the representation (8.9) is of major interest because we can use standard simulation procedures for its approximation. First, for a given value of β, $\mathbb{E}_\beta[S(\mathbf{X})]$ can be approximated from an MCMC sequence simulated by Algorithm 8.3. Obviously, changing the value of β requires a new simulation run, but the cost can be attenuated by using importance sampling for similar values of

β. Second, the integral itself can be approximated by *numerical quadrature,* namely by computing the value of $f(\beta) = \mathbb{E}_\beta[S(\mathbf{X})]$ for a finite number of values of β and approximating $f(\beta)$ by a piecewise-linear function $\hat{f}(\beta)$ for the intermediate values of β. Indeed,

$$\int_{\beta_0}^{\beta_1} \hat{f}(\beta)\,\mathrm{d}\beta = f(\beta_0) + \{f(\beta_1) - f(\beta_0)\}\,\frac{(\beta_1 - \beta_0)^2}{2}$$

if β_0 and β_1 are two consecutive values, where $f(\beta)$ is approximated by Monte Carlo methods.

In the Menteith setting, we have a four-neighbor neighborhood and $G = 6$ on a 100×100 image. For β ranging from 0 to 2 by steps of 0.1, the approximation to $f(\beta)$ is based on 1500 iterations of Algorithm 8.3 (after burn-in). The piecewise-linear function is given in Figure 8.10 and is smooth enough for us to consider the approximation is valid. Note that the increasing nature of the function f in β is intuitive: As β grows, the probability of having more neighbors of the same category increases and so does $S(\mathbf{x})$.

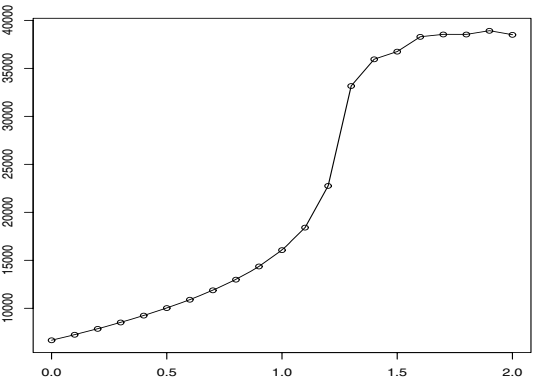


Fig. 8.10. Approximation of $f(\beta)$ for the Potts model on a 100×100 image, a four-neighbor neighborhood, and $G = 6$, based on 1500 MCMC iterations after burn-in.

Exercise 8.15. Using a piecewise-linear interpolation of $f(\beta)$ based on the values $f(\beta^1), \ldots, f(\beta^M)$, with $0 < \beta_1 < \ldots < \beta_M = 2$, give the explicit value of the integral

$$\int_{\alpha_0}^{\alpha_1} \hat{f}(\beta)\,\mathrm{d}\beta$$

for any pair $0 \leq \alpha_0 < \alpha_1 \leq 2$.

Ⓒ Now that we have painstakingly constructed a satisfactory approximation of $Z(\beta_1)/Z(\beta_0)$ for any arbitrary pair of β's, we can run a hybrid Gibbs sampler corresponding to the full distribution $\pi(\mathbf{x}, \beta, \sigma^2, \boldsymbol{\mu}|\mathbf{y})$, where simulation for β at iteration t is based on the proposal

$$\tilde{\beta} \sim \mathcal{U}([\beta^{(t-1)} - h, \beta^{(t-1)} + h])\,;$$

that is, a uniform move with range $2h$. The acceptance ratio is thus given by

$$1 \wedge \left(\widehat{Z(\beta^{(t-1)})/Z(\tilde{\beta})}\right) \exp\left\{(\tilde{\beta} - \beta^{(t-1)})S(\mathbf{x})\right\}\,.$$

⊞ Figures 8.11 to 8.13 illustrate the convergence performances of the hybrid Gibbs sampler for **Menteith**. In that case, using $h = 0.05$ shows that 2000 MCMC iterations are sufficient for convergence. (Recall, however, that $\mathbf{x}$ is a 100×100 image and thus that a single Gibbs step implies simulating the value of 10^4 pixels. This comes in addition to the cost of approximating the ratio of normalizing constants.) All histograms are smooth and unimodal, even though the moves on β are more difficult than for the other components. Different values of h were tested for this dataset and none improved this behavior. Note that large images like **Menteith** often lead to a very concentrated posterior on β. (Other starting values for β were also tested to check for the stability of the stationary region.)

Remember that the primary purpose of this analysis is to clean (de-noise) and to classify into G categories the pixels of the image. Based on the MCMC output and in particular on the chain $(\mathbf{x}^{(t)})_{1\le t\le T}$ (where T is the number of MCMC iterations), an estimator of $\mathbf{x}$ needs to be derived from an evaluation of the consequences of wrong allocations. Two common ways of running this evaluation are either to count the number of (individual) pixel misclassifications,

$$L_1(\mathbf{x}, \widehat{\mathbf{x}}) = \sum_{i\in\mathcal{I}} \mathbb{I}_{x_i \neq \hat{x}_i}\,,$$

or to use the global "zero–one" loss function (see Section 2.3.1),

$$L_2(\mathbf{x}, \widehat{\mathbf{x}}) = \mathbb{I}_{\mathbf{x}\neq\widehat{\mathbf{x}}}\,,$$

which amounts to saying that only a perfect reconstitution of the image is acceptable (and thus sounds rather extreme in its requirements). It is then easy to show that the estimators associated with both loss functions are the marginal posterior mode (MPM), $\widehat{\mathbf{x}}^{MPM}$; that is, the image made of the pixels

$$\hat{x}_i^{MPM} = \arg\max_{1\le g\le G} \mathbb{P}^{\pi}(x_i = g|\mathbf{y})\,, \quad i \in \mathcal{I}\,,$$

and the maximum a posteriori estimator (2.4),

$$\widehat{\mathbf{x}}^{MAP} = \arg\max_{\mathbf{x}} \pi(\mathbf{x}|\mathbf{y})\,,$$

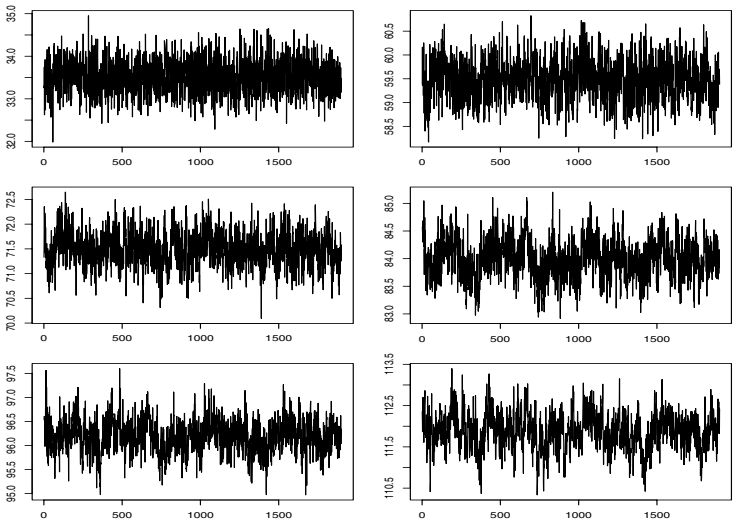


Fig. 8.11. Dataset Menteith: Sequence of μ_g's based on 2000 iterations of the hybrid Gibbs sampler (*read row-wise from* μ_1 *to* μ_6).

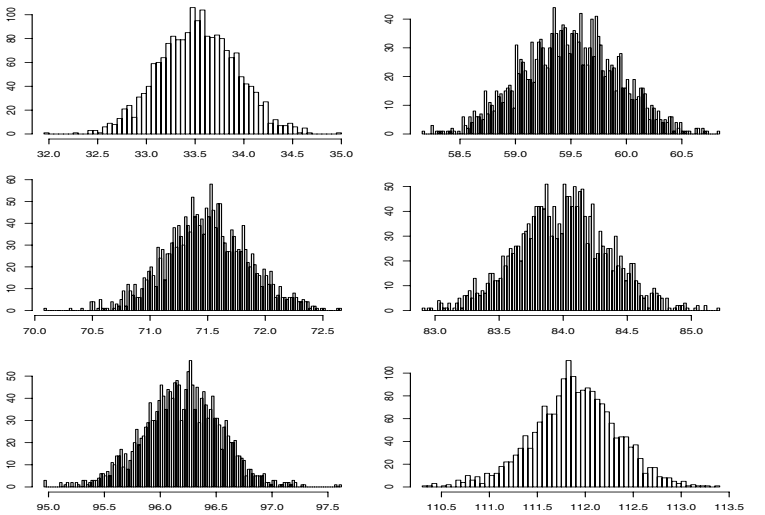


Fig. 8.12. Dataset Menteith: Histograms of the μ_g's represented in Figure 8.11.

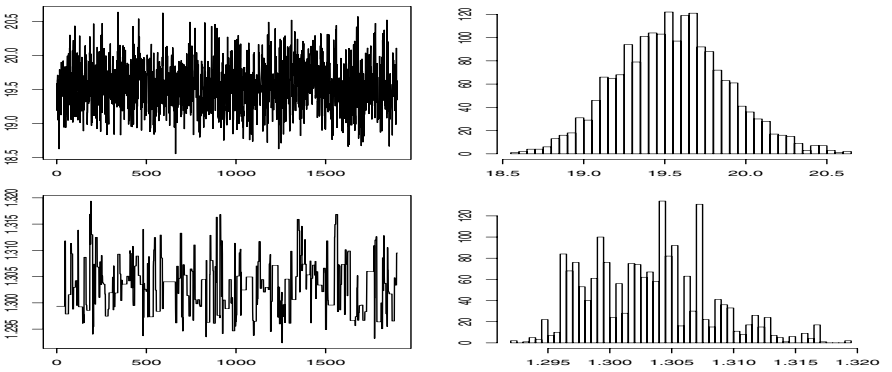


Fig. 8.13. Dataset Menteith: Raw plots and histograms of the σ^2's and β's based on 2000 iterations of the hybrid Gibbs sampler (*the first row corresponds to* σ^2).

respectively. Note that it makes sense that the $\widehat{\mathbf{x}}^{MPM}$ estimator only depends on the marginal distribution of the pixels, given the linearity of the loss function. Both loss functions are nonetheless associated with image reconstruction rather than true classification (Exercise 8.17).

Exercise 8.16. Show that the estimators $\widehat{\mathbf{x}}$ that minimize the posterior expected losses $\mathbb{E}[L_1(\mathbf{x},\widehat{\mathbf{x}})|\mathbf{y})]$ and $\mathbb{E}[L_2(\mathbf{x},\widehat{\mathbf{x}})|\mathbf{y}]$ are $\widehat{\mathbf{x}}^{MPM}$ and $\widehat{\mathbf{x}}^{MAP}$, respectively.

Exercise 8.17. Determine the estimators $\widehat{\mathbf{x}}$ associated with two loss functions that penalize differently the classification errors,

$$L_3(\mathbf{x},\widehat{\mathbf{x}}) = \sum_{i,j\in\mathcal{I}} \mathbb{I}_{x_i=x_j}\,\mathbb{I}_{\hat{x}_i\neq\hat{x}_j} \quad \text{and} \quad L_4(\mathbf{x},\widehat{\mathbf{x}}) = \sum_{i,j\in\mathcal{I}} \mathbb{I}_{x_i\neq x_j}\,\mathbb{I}_{\hat{x}_i=\hat{x}_j}\,.$$

Ⓒ The estimators $\widehat{\mathbf{x}}^{MPM}$ and $\widehat{\mathbf{x}}^{MAP}$ obviously have to be approximated since the marginal posterior distributions $\pi(x_i|\mathbf{y})$ $(i\in\mathcal{I})$ and $\pi(\mathbf{x}|\mathbf{y})$ are not available in closed form. The marginal distributions of the x_i's being by-products of the MCMC simulation of $\mathbf{x}$, we can use, for instance, as an approximation to $\widehat{\mathbf{x}}^{MPM}$ the most frequent occurrence of each pixel $i\in\mathcal{I}$,

$$\hat{x}_i^{MPM} = \max_{g\in\{1,\ldots,G\}} \sum_{j=1}^{N} \mathbb{I}_{x_i^{(j)}=g}\,,$$

based on a simulated sequence, $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)}$, from the posterior distribution of $\mathbf{x}$. (This is not the most efficient approximation to $\widehat{\mathbf{x}}^{MPM}$, obviously, but it comes as a cheap by-product of the MCMC simulation and it does not require the use of more advanced simulated annealing tools, mentioned in Section 6.6.)

Unfortunately, the same remark cannot be made about $\widehat{\mathbf{x}}^{MAP}$: The state space of the simulated chain $(\mathbf{x}^{(t)})_{1\le t\le T}$ is so huge, being of cardinality $G^{100\times 100}$, that it is com-

pletely unrealistic to look for a proper MAP estimate out of the sequence $(\mathbf{x}^{(t)})_{1\le t\le T}$. Since $\pi(\mathbf{x}|\mathbf{y})$ is not available in closed form, even though this density could be approximated by

$$\widehat{\pi}(\mathbf{x}|\mathbf{y}) \propto \sum_{t=1}^{T} \pi(\mathbf{x}|\mathbf{y}, \beta^{(t)}, \boldsymbol{\mu}^{(t)}, \sigma^{(t)}) \,,$$

thanks to a Rao–Blackwellization argument, it is rather difficult to propose a foolproof simulated annealing that converges to $\widehat{\mathbf{x}}^{MAP}$ (although there exist cheap approximations; see Exercise 8.18).

Exercise 8.18. Since the maximum of $\pi(\mathbf{x}|\mathbf{y})$ is the same as that of $\pi(\mathbf{x}|\mathbf{y})^\kappa$ for every $\kappa \in \mathbb{N}$, show that

$$\pi(\mathbf{x}|\mathbf{y})^\kappa = \int \pi(\mathbf{x}, \theta_1|\mathbf{y}), \mathrm{d}\theta_1 \times \cdots \times \int \pi(\mathbf{x}, \theta_\kappa|\mathbf{y}), \mathrm{d}\theta_\kappa \,, \tag{8.10}$$

where $\theta_i = (\beta_i, \boldsymbol{\mu}_i, \sigma_i^2)$ $(1 \le i \le \kappa)$. Deduce from this representation an optimization scheme that slowly increases κ over iterations and that runs a Gibbs sampler for the integrand of (8.10) at each iteration.

⊞ The segmented image of Lake Menteith is given by the MPM estimate that was found after 2000 iterations of the Gibbs sampler. We reproduce in Figure 8.14 the original picture to give an impression of the considerable improvement brought by the algorithm.

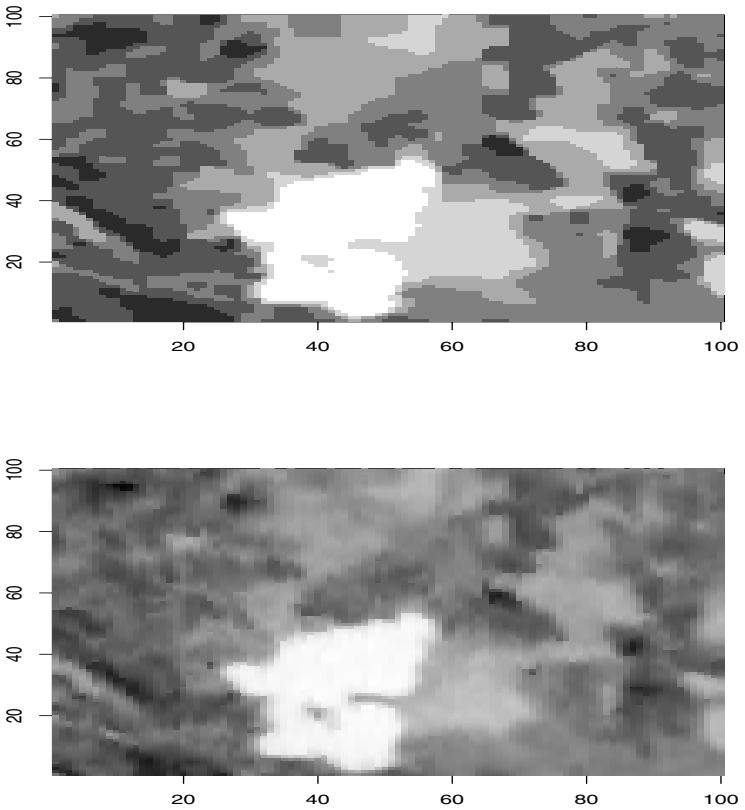


Fig. 8.14. Dataset Menteith: (*top*) Segmented image based on the MPM estimate produced after 2,000 iterations of the Gibbs sampler and (*bottom*) the observed image.

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley, New York.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley, New York.

Brockwell, P. and Davis, P. (1996). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer-Verlag, New York.

Cappé, O., Moulines, E., and Rydén, T. (2004). *Hidden Markov Models*. Springer-Verlag, New York.

Carlin, B. and Louis, T. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York.

Casella, G. and Berger, R. (2001). *Statistical Inference*. Wadsworth, Belmont, CA.

Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P. (1983). *Graphical Methods for Data Analysis*. Chapman and Hall, New York.

Chen, M., Shao, Q., and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

Chown, M. (1996). *Afterglow of Creation: From the Fireball to the Discovery of Cosmic Ripples*. University Science Books, Mill Valley, CA.

Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer Texts in Statistics. Springer-Verlag, New York.

Congdon, P. (2001). *Bayesian Statistical Modelling*. John Wiley, New York.

Congdon, P. (2003). *Applied Bayesian Modelling*. John Wiley, New York.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley, New York.

Dalgaard, P. (2002). *Introductory Statistics with R*. Springer-Verlag, New York.

Dupuis, J. (1995). Bayesian estimation of movement probabilities in open populations using hidden Markov chains. *Biometrika*, 82(4):761–772.

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics, A Practical Approach*. Cambridge University Press, Cambridge.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2001). *Bayesian Data Analysis*. Chapman and Hall, New York, second edition.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741.

Gentle, J. (2002). *Elements of Computational Statistics*. Springer-Verlag, New York.

Gill, J. (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach*. CRC Press, Boca Raton, FL.

Gouriéroux, C. (1996). *ARCH Models*. Springer-Verlag, New York.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Holmes, C., Denison, D., Mallick, B., and Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley, New York.

Hurn, M., Husby, O., and Rue, H. (2003). A Tutorial on Image Analysis. In Møller, J., editor, *Spatial Statistics and Computational Methods*, volume 173 of *Lecture Notes in Statistics*, pages 87–141. Springer-Verlag, New York.

Lebreton, J.-D., Burnham, K., Clobert, J., and Anderson, D. (1992). Modelling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, 62(2):67–118.

Lee, P. (1997). *Bayesian Statistics: An Introduction*. Oxford University Press, Oxford, second edition.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, New York.

McDonald, I. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall/CRC, London.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, New York.

Møller, J. (2003). *Spatial Statistics and Computational Methods*, volume 173 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Nolan, D. and Speed, T. (2000). *Stat Labs: Mathematical Statistics through Applications*. Springer-Verlag, New York.

Pole, A., West, M., and Harrison, J. (1994). *Applied Bayesian Forecasting and Time Series Analysis*. Chapman and Hall, New York.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Soc. Series B*, 59:731–792.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, second edition.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition.

Särndal, C., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer-Verlag, New York, second edition.

Schervish, M. (1995). *Theory of Statistics*. Springer-Verlag, New York.

Tanner, M. (1996). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. Springer-Verlag, New York, third edition.

Tomassone, R., Dervin, C., and Masson, J.-P. (1993). *Biométrie : Modélisation de Phénomènes Biologiques*. Masson, Paris.

Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer-Verlag, New York, fourth edition.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley, Chichester.

Zellner, A. (1971). *An Introduction to Bayesian Econometrics*. John Wiley, New York.

Zellner, A. (1984). *Basic Issues in Econometrics*. University of Chicago Press, Chicago.

Index

Springer Texts in Statistics *(continued from p. ii)*

Robert/Casella: Monte Carlo Statistical Methods, Second Edition
Rose/Smith: Mathematical Statistics with *Mathematica*
Ruppert: Statistics and Finance: An Introduction
Sen/Srivastava: Regression Analysis: Theory, Methods, and Applications
Shao: Mathematical Statistics, Second Edition
Shorack: Probability for Statisticians
Shumway/Stoffer: Time Series Analysis and Its Applications, Second Edition
Simonoff: Analyzing Categorical Data
Terrell: Mathematical Statistics: A Unified Introduction
Timm: Applied Multivariate Analysis
Toutenberg: Statistical Analysis of Designed Experiments, Second Edition
Wasserman: All of Nonparametric Statistics
Wasserman: All of Statistics: A Concise Course in Statistical Inference
Weiss: Modeling Longitudinal Data
Whittle: Probability via Expectation, Fourth Edition