Use R!

Ingmar Visser
Maarten Speekenbrink

Mixture and Hidden Markov Models with R

Springer

Use R!

This series of inexpensive and focused books on R is aimed at practitioners. Books can discuss the use of R in a particular subject area (e.g., epidemiology, econometrics, psychometrics) or as it relates to statistical topics (e.g., missing data, longitudinal data). In most cases, books combine LaTeX and R so that the code for figures and tables can be put on a website. Authors should assume a background as supplied by Dalgaard's Introductory Statistics with R or other introductory books so that each book does not repeat basic material.

How to Submit Your Proposal

Book proposals and manuscripts should be submitted to one of the publishing editors in your region per email – for the list of statistics editors by their location please see https://www.springer.com/gp/statistics/contact-us. All submissions should include a completed Book Proposal Form.

For general and technical questions regarding the series and the submission process please contact Laura Briskman (laura.briskman@springer.com) or Veronika Rosteck (veronika.rosteck@springer.com).

Ingmar Visser • Maarten Speekenbrink

Mixture and Hidden Markov Models with R

Springer

Ingmar Visser
University of Amsterdam
Amsterdam, The Netherlands

Maarten Speekenbrink
University College London
London, UK

ISSN 2197-5736 ISSN 2197-5744 (electronic)
Use R!
ISBN 978-3-031-01438-3 ISBN 978-3-031-01440-6 (eBook)
https://doi.org/10.1007/978-3-031-01440-6

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Lieke and Pien
I.V.

To Gabriel and Hunter
M.S.

Preface

This book aims to provide a self-contained practical introduction to mixture models and hidden Markov models. The reason for introducing both in one book is that there are very close links between these models. This allows us to introduce important concepts, such as maximum likelihood estimation and the Expectation-Maximization algorithm, in the relatively simpler context of mixture models. Approaching hidden Markov models from a thorough understanding of mixture models involves, we hope, a relatively small conceptual leap. We aimed to provide a reasonable balance between statistical theory and practice. The objective is to provide enough mathematical details—but no more!—to allow our target audience to understand key results that are necessary to apply these models.

Our target audience are those with a more applied background, in particular researchers, graduate, and advanced undergraduate students in the social and behavioral sciences. Researchers or future researchers hence who see the potential for applying these models and explaining heterogeneity in their data, but who currently lack the tools to fulfill this potential. Those looking for a more purely mathematical treatment of mixture and hidden Markov models we gladly refer to the books by Cappe et al. (2005) and/or Frühwirth-Schnatter (2006).

To familiarize readers with the possibilities of mixture and hidden Markov models, a large part of this book consists of practical examples of applying these models, many of which taken from our own research in developmental and experimental psychology. Much of our work on these models was driven by the research questions that arose during the study of experimental or developmental (time series) data. Over the years, we have also accumulated examples from other fields, such as climate change and economics. In the examples, we provide some background knowledge of these different domains as applicable to understand the rationale of the analyses. At the same time, we abstract away from many details and focus on the generalizability of the presented models to research questions in other domains.

The example analyses in this book rely on the R programming language and software environment (R Core Team 2020) and in particular the **depmixS4** package (Visser and Speekenbrink 2010). Nowadays, the choice for R hardly needs justification, having become the lingua franca of statistics and data science. R is open source, freely available, and has an active user community such that anyone interested can add and contribute packages implementing new analytical methods. As all required tools are freely available, the readers should be able to replicate the example analyses on their own computers, as well as adapting the analyses for their own purposes. To aid in this process, all the code for running the examples in the book are provided online at https://depmix.github.io/hmmr/. Moreover, the datasets and special purpose functions written for this book are available as an R package called **hmmr**. Section 1.2 provides pointers for getting started with R and provides all the basics that are needed to then understand and apply subsequent analyses and examples.

Chapter Outlines and Reading Guide

Chapter 1 provides a brief introduction to R and describes the basic features of the datasets used throughout this book to illustrate the use of mixture and hidden Markov models. Chapters 2 and 4 are mostly theoretical in nature, providing a statistical treatment of mixture and latent class models (Chap. 2), and the extension of those models into hidden Markov models (Chap. 4). Chapter 3 provides a number of worked examples of applications of mixture and latent class models to analyze both univariate and multivariate data. Similarly, Chaps. 5 and 6 provide detailed example analyses which apply hidden Markov models to univariate (Chap. 5) and multivariate (Chap. 6) time series data. Finally, Chap. 7 discusses some extensions of the basic hidden Markov model, as well as alternative estimation techniques, including a brief introduction to Bayesian estimation of these models.

In Chaps. 2 and 4, the first two sections are devoted to conceptually describing and defining mixture and hidden Markov models, respectively. These sections lay the foundations for understanding how these models work and how they can be usefully applied. These sections should be read by everyone. Rushed readers wanting to get started right away with applying the models may skip the remainder of those chapters, where we delve deeper into parameter estimation and inference. The examples in Chaps. 3, 5, and 6, are standalone sections that treat data with particular characteristics and describe the models that can be used to answer the research questions of interest. Where warranted, these application sections also refer back to the relevant sections in Chaps. 2 and 4 which offer more technical detail of topics that arise. Readers who skipped most of Chaps. 2 and 4 can then read the relevant parts of these chapters when the need arises.

Acknowledgments

Writing and producing a book is rarely done in isolation, and this one is no exception. Many people have contributed by asking tough research questions, providing data, LaTeX, and `Sweave()` advice. Below is a list of people that we know have surely contributed in important ways. We know that this list is likely incomplete, so just let us know if you ought to be on the list so we can include you in future editions. We would like to thank Achim Zeileis for getting us started with the combination of LaTeX, `Sweave()` and make files to produce this book. We would like to thank Chen Haibo for providing the S&P-500 data example, Brenda Jansen for sharing her balance scale data, Gilles Dutilh and Han van der Maas for sharing their speed-accuracy data, the Water Corporation of Western Australia for providing the Perth dams data on their website, Bianca van Bers for sharing the dimensional change card sorting task data, Han van der Maas for sharing the conservation of liquid data, Maartje Raijmakers for sharing the discrimination data, and Emmanouil Konstantinidis for sharing the Iowa gambling task data. Finally, we would like to thank John Kimmel, Marc Strauss, and Laura Briskman for inviting us to write this book and organize things Springer.

This book has been taking us a while to complete. On a more personal level, many people have been by our sides during that period. Maarten would like to thank Gabriel and Hunter for being wondering and wonderful human beings, and Ria and Jan for their love and support. Ingmar would like to thank Jaro for her love and support.

Amsterdam, The Netherlands — Ingmar Visser
London, UK — Maarten Speekenbrink
January 2022

Settings, Appearance, and Notation

In producing the examples in this book, R is mainly run at its default settings. A few modifications were made to render the output more easily readable; these were invoked by the following chunk of R-code:

```
R> options(prompt = "R> ", continue = "+  ", width = 60,
+    digits = 4, show.signif.stars = FALSE,
+    useFancyQuotes = FALSE)
```

This replaces the standard R prompt `>` by `R>`. For compactness, `digits = 4` reduces the number of digits shown when printing numbers from the default of 7. Note that this does not reduce the precision with which these numbers are internally processed and stored.

We use `set.seed(x)` whenever we generate data or fit models such that the exact values of data and fitted parameters may be replicated. When fitting models, this is necessary, because random starting values are generated (see Sect. 2.3.6 for more details).

We use a `typewriter` font for all code; additionally, function names are followed by parentheses, as in `plot()`, and class names (a concept that is explained in Chap. 1) are displayed as in "`depmix`." Furthermore, boldface is used for package names, as in **hmmr**.

The following symbols are used throughout the book:

$\mathbf{A}$	Transition matrix
π	Initial state probability vector
S	Stochastic state variable
s	Realization of the state variable
Y	Stochastic (possibly multivariate) observation variable
y	Realization of the observation variable
z	Covariate, possibly multivariate
$\boldsymbol{\theta}$	Total model parameter vector; $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{pr}}, \boldsymbol{\theta}_{\text{tr}}, \boldsymbol{\theta}_{\text{obs}})$
$\boldsymbol{\theta}_{\text{pr}}$	Subvector of the parameter vector with parameters of the prior model
$\boldsymbol{\theta}_{\text{tr}}$	Subvector of the parameter vector with parameters of the transition model

$\boldsymbol{\theta}_{\text{obs}}$	Subvector of the parameter vector with parameters of the observation model
T	Total number of time points
N	Number of states of a model
f	Probability density function
P	Probability distribution
$\mathbf{H}$	Hessian matrix

Contents