Latent Markov Models for Longitudinal Data

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Aims and scope

Large and complex datasets are becoming prevalent in the social and behavioral sciences and statistical methods are crucial for the analysis and interpretation of such data. This series aims to capture new developments in statistical methodology with particular relevance to applications in the social and behavioral sciences. It seeks to promote appropriate use of statistical, econometric and psychometric methods in these applied sciences by publishing a broad range of reference works, textbooks and handbooks.

The scope of the series is wide, including applications of statistical methodology in sociology, psychology, economics, education, marketing research, political science, criminology, public policy, demography, survey methodology and official statistics. The titles included in the series are designed to appeal to applied statisticians, as well as students, researchers and practitioners from the above disciplines. The inclusion of real examples and case studies is therefore essential.

Published Titles

Analysis of Multivariate Social Science Data, Second Edition
David J. Bartholomew, Fiona Steele, Irini Moustaki, and Jane I. Galbraith

Applied Survey Data Analysis
Steven G. Heeringa, Brady T. West, and Patricia A. Berglund

Bayesian Methods: A Social and Behavioral Sciences Approach, Second Edition
Jeff Gill

Foundations of Factor Analysis, Second Edition
Stanley A. Mulaik

Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists
Herbert Hoijtink

Latent Markov Models for Longitudinal Data
Francesco Bartolucci, Alessio Farcomeni, and Fulvia Pennoni

Linear Causal Modeling with Structural Equations
Stanley A. Mulaik

Multiple Correspondence Analysis and Related Methods
Michael Greenacre and Jorg Blasius

Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences
Brian S. Everitt

Statistical Test Theory for the Behavioral Sciences
Dato N. M. de Gruijter and Leo J. Th. van der Kamp

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Latent Markov Models for Longitudinal Data

Francesco Bartolucci
Alessio Farcomeni
Fulvia Pennoni

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20150813

International Standard Book Number-13: 978-1-4665-8371-9 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

List of Figures

List of Tables

Preface

Latent Markov models represent an important class of latent variable models for the analysis of longitudinal data, when the response variables measure common characteristics of interest which are not directly observable. Typically, the response variables are categorical, even if nothing precludes that they have a different nature. These models find application in many relevant fields, such as educational and health sciences, when the latent characteristics correspond, for instance, to a certain type of ability or to the quality of life. Important applications are also in the study of certain human behaviors which are relevant for social and economic research.

The main feature that distinguishes latent Markov models from other models for longitudinal data is that the individual characteristics of interest, and their evolution in time, are represented by a latent process which follows a Markov chain. This implies that we are in the field of discrete latent variable models, where the latent variables may assume a finite number of values. Latent Markov models are then strongly related to the latent class model, which is an important tool for classifying a sample of subjects on the basis of a series of categorical response variables. The latter model is based on a discrete latent variable, the different values of which correspond to different subpopulations (named *latent classes*) having a common distribution about the response variables. The latent Markov model may be seen as an extension of the latent class model in which subjects are allowed to move between the latent classes during the period of observation.

Latent Markov models are also naturally related to the Markov chain model. In applications involving longitudinal data, the latter is typically used to study the transition between observable states, possibly depending on individual covariates. In particular, a latent Markov model may be seen as an extension, which accounts for measurement errors, of the Markov chain model. A strong connection also exists with hidden Markov models. In particular, latent Markov and hidden Markov models share the same assumptions and estimation methods. The main difference is in the structure of the data that they are aimed at analyzing. Latent Markov models are typically used in the context of longitudinal data, whereas hidden Markov models are used for the analysis of time-series data. Roughly speaking, in the first case we observe short sequences of data in correspondence to a large number of individuals or statistical units, whereas in the second case we observe long sequences of data referred to one or few statistical units. This difference has implications on the foundation of the asymptotic theory about the properties of the estimators for these models. In the context of longitudinal data, asymptotic properties are proved under the assumption that the sample size tends to infinity. In the context of time series, asymptotic properties postulate that the number of occasions of observation tends to infinity. However, the terminology is not univocal and the name *hidden Markov model* is sometimes adopted even for models applied to the analysis of longitudinal data.

Through the present book, we aim at providing the reader with a complete overview of latent Markov models, with special attention to the interpretation of the model assumptions and to the practical use of these models. We begin with an outline, given in Chapter 1, of the role of these models for modern applications and of the main bibliography on the topic. The same chapter describes some datasets, which cover fields of economics, education, and sociology and that will be used to illustrate the proposed approaches. Then, Chapter 2 provides the reader with the essential background about latent variable models, and in particular the latent class model. We also review the Markov chain model that, together with latent class model, represents a useful paradigm for latent Markov models. Chapter 3 illustrates the assumptions of the basic version of the latent Markov model, which is used in the presence of univariate or multivariate responses, but without covariates. This chapter also introduces maximum likelihood estimation through the Expectation-Maximization algorithm. Then, Chapter 4 introduces some constrained versions of the basic latent Markov model based on parsimonious and interpretable parametrizations. Chapter 5 discusses the inclusion of the individual covariates, whereas Chapter 6 introduces the random-effects and the multilevel extensions of the model. Chapter 7 covers some more advanced topics, such as the performance of criteria for selecting the number of latent states and path prediction. Chapter 8 introduces Bayesian inference as an alternative to maximum likelihood estimation. Some of the examples that we will use to illustrate the content of the book chapters are developed by using an R package, named `LMest`, which we make available to the reader[1] and that is described in the final Appendix. Other examples have been based on other R and `MATLAB` routines which are available upon request.

[1] At the website `www.stat.unipg.it/bartolucci`

To easily learn the content of the present book, a modest knowledge of probability and statistics is required. In particular we expect the reader to know the basic concepts about maximum likelihood estimation and Bayesian theory. The reader may wish to supplement our treatment by referring to some relevant parts of general textbooks of statistics, such as Casella and Berger (2002), and to books dealing in particular with the maximum likelihood approach, such as Azzalini (1996). Concerning Chapter 8 about Bayesian inference, some prior knowledge on Markov chain Monte Carlo algorithms is expected and, in particular, about the Reversible Jump algorithm. In this regard, readers would find profitable to refer to Robert and Casella (2010) or Ghosh et al. (2010). The present book is related to other books which develop the theory of hidden Markov models, such as Elliott et al. (1995), MacDonald and Zucchini (1997), Cappé et al. (2005), Zucchini and MacDonald (2009), and Dymarski (2011).

Finally, we have to mention that the book is the result of the intense research activity of the authors in the field of the analysis of categorical longitudinal data, which has been developed in the last years. We acknowledge the financial support to this activity provided by the Italian Government (research grant PRIN `2007XECZ7L_003`) and, in particular, by the "Einaudi Institute for Economics and Finance." We are also grateful to Silvia Pandolfi, of the Department of Economics, Finance, and Statistics of the University of Perugia (IT), for her relevant aid in writing Chapter 8 about Bayesian inference and for the careful reading of all the material.

1

Overview on latent Markov modeling

1.1 Introduction

Modern applications in the field of sociology and economics, and in many other fields, are frequently based on longitudinal (or panel) data. These data arise from the repeated observation of the same subjects, or more generally sample units, at a certain number of time occasions. From many points of view, longitudinal data are similar to time-series data, even if the contexts of application are usually very different. The main difference is that time series usually refer to only one unit observed at many time occasions, whereas longitudinal data refer to several units observed at few time occasions.

The increasing relevance of longitudinal data is witnessed by the widespread use of these data in many research fields. An important example is the data collected within the Panel Study of Income Dynamics (PSID), which are freely accessible.[1] Another example of freely available longitudinal data is those collected within the Health and Retirement Study conducted by the University of Michigan.[2] Many other interesting longitudinal data are freely provided by the UK Longitudinal Study Centre.[3] For a deeper discussion of these and similar databases see Hsiao (2003), Chapter 1.

The main advantage of using longitudinal data is that they allow us to understand the evolution of a certain behavior or phenomenon across time. Moreover, with respect to cross-sectional data, arising from the observation of the subjects at only one occasion, longitudinal data allow for a more precise assessment of the effect of covariates (or explanatory variables) on the response variables. This is because the same unit is observed under different circumstances corresponding to different configurations of the explanatory variables. This aspect is particularly relevant when these variables are related to a certain treatment or policy. The

[1] See the website `http://psidonline.isr.umich.edu/`

[2] A detailed description of the data can be found at the following website: `http://www.rand.org/labor/aging/dataprod`

[3] See the website `http://www.esds.ac.uk/longitudinal/about/introduction.asp`

causal effect of treatment may be more easily measured than using cross-sectional data. On the other hand, longitudinal data frequently present missing observations, especially when the sample is followed for a long period of time. This is due to subjects leaving the panel under observation. Ignoring these units may induce a strong bias in the conclusions of the study. In any case, the degree of informativeness of longitudinal data is obviously much higher than that of cross-sectional data collected on a sample of subjects of the same dimension.

A great variety of statistical and econometric models now exists for the analysis of longitudinal data; for a review see, among others, Diggle et al. (2002), Frees (2004), Hsiao (2003), and Fitzmaurice et al. (2009). The choice of the model depends on the specific context of application and on the nature of the response and explanatory variables of interest. In this framework, latent Markov (LM) models, which are a particular family of models for longitudinal data and are the main subject of this book, play a special role.

As we will discuss in more detail in the following, the basic formulation of LM models for longitudinal data is the same as hidden Markov (HM) models, which have been earlier developed in the literature on stochastic processes and on time-series data; for an up-to-date review see Zucchini and MacDonald (2009) and Dymarski (2011). However, the point of view is rather different in light of the type of data which are encountered: time series are usually made up of many repeated measures referred to a single unit, whereas only a few repeated measures are typically available (but for many sample units) in a longitudinal dataset.

LM models assume the existence of a latent process which affects the distribution of the response variables. We may have one response variable for each time occasion, and then *univariate longitudinal data*, or more than one response variable for each time occasion, and then *multivariate longitudinal data*. The latent process is assumed to follow a Markov chain with a certain number of states, typically referred to as *latent states*, and, given this process, the response variables are assumed to be conditionally independent. The latter assumption is known as *local independence*. In this regard, it is important to distinguish between two components of an LM model: the *measurement model* and the *latent model*. They concern the distribution of the response variables given the latent process (in the first case) and the distribution of the latent process (in the second case). Individual covariates may be naturally included in this framework and assumed to enter either in the measurement model or in the latent model.

In this book, we provide a review of LM models with reference to the case of longitudinal data with categorical response variables, even if we briefly discuss the case of response variables of a different nature. In the longitudinal data context, the use of the LM framework finds justification in different types of analysis, which we now outline. For each type of analysis we briefly introduce an example that will be later developed.

1. *Transition analysis with measurement errors.* This is typical of applications based on univariate longitudinal data when the "true state" characterizing a sample unit at a certain occasion may be measured with error. The "true states" correspond to the latent states and transition between each pair of these states is studied through the transition probabilities. Moreover, if they exist, individual covariates are typically assumed to affect the latent model. It is rather obvious that, under this interpretation, the adopted LM model is seen as an extension of a Markov chain (MC) model (Anderson, 1951, 1954). Then, by comparing an LM model with an MC model we can test for the absence of measurement errors. As an example, consider the case in which the response variable is the self-reported level of use of a certain drug. In this case, we can interpret the latent states as true levels of drug use which are reported with possible error by the respondent. Then, the transition probabilities say how subjects evolve in the true drug use level, whereas the conditional response probabilities say how the reported drug use level depends on the true one.

2. *Analyses which take into account unobserved heterogeneity.* Unobserved heterogeneity may be defined as the heterogeneity between the responses provided by different individuals that cannot be explained on the basis of observable individual covariates. This is an aspect which should often be taken into account in the analysis of longitudinal data, both univariate and multivariate. Through an LM formulation, in which the latent states are associated to different levels of the effect of the unobservable covariates on the response variables, we model the unobserved heterogeneity in a dynamic fashion because we admit that each individual may belong to different latent states during the period of observation. In other terms, we admit that each subject may move between these latent states. Note that the way in which we take the unobserved heterogeneity into account is by discrete latent variables, which is less common than the approach based on continuous latent variables (or *random effects*). The discrete latent variable approach is typical of the latent class (LC) model (Lazarsfeld, 1950; Lazarsfeld and Henry, 1968). However, the LM approach generalizes the LC approach by allowing subjects to move between latent states, whereas under the LC model every subject always remains in the same latent state, which is known as *latent class.* Note that, in this

context, even an LM model without covariates makes sense since it may be used as a counterpart against which the LC model may be tested. On the other hand, if observable covariates exist, they are included in the measurement model. As an example consider the case in which we observe, at repeated occasions, a binary response variable indicating if an individual has a job position or not. Then, through an LM model we can take into account the effect of unobservable factors, such as motivations or intelligence, which affect this variable.

3. *Finding clusters of units and studying the transition between these clusters.* This type of analysis is typically based on multivariate longitudinal data, which is when more response variables are observed at each occasion. Through an LM model suitably formulated, we can aggregate subjects in different clusters, corresponding to the latent states, in a way similar to an LC model. However, we have a higher degree of flexibility due to the possibility of subjects of moving between these clusters. Obviously, even the way in which subjects move between the clusters may be of interest and may be explained on the basis of transition probabilities; in this sense we are performing an analysis similar to that aimed at studying transitions with measurement errors (see item 1 above). If available, covariates are included in the latent model and then may affect initial and transition probabilities of the Markov chain. As an example, consider the case of the analysis of criminal data in which for each age band (period of time) we know which types of crime are committed by subjects in a certain cohort. We can first classify subjects in different clusters and then study the transition between different clusters. Each cluster corresponds to a latent state, with subjects in the same latent state having the same criminal behavior.

1.2 Literature review on latent Markov models

As already mentioned, LM models are strongly related to HM models for time-series and stochastic processes; for general reviews see the monographs by MacDonald and Zucchini (1997), Koski (2001), and Zucchini and MacDonald (2009), whereas for a review about estimation methods for these models see Cappé et al. (2005). Both HM and LM approaches rely on a latent process given which the response variables are conditionally independent. This process is assumed to follow a Markov chain, typically of first order. For this reason, LM models for longitudinal data are sometimes called HM models. However, in the present book we specifically use the terminology HM in the context of time-series data and LM in the context of longitudinal data.

Even if LM and HM models share the same basic assumptions, the specific developments have been carried out in separated fields of the statistical literature and related literatures. In fact, time-series and longitudinal data present specific issues. Just to give an example, in the context of time-series data analysis, asymptotic properties of an estimator are studied assuming that the number of repeated measures grows to infinity, whereas in the context of longitudinal data analysis, asymptotic properties are studied assuming that the sample size tends to infinity.

The literature on HM models appears to have been developed earlier than that on LM models, as many important results have been elaborated in the 60's and 70's. In fact, one of the most relevant papers is that by Baum and Petrie (1966) concerning inference for basic HM models. Moreover, many of these results have been elaborated in connection with engineering, informatics, and bioinformatics applications and then have even appeared in papers published in nonstatistical journals; consider, for instance, the papers by Levinson et al. (1983) and Ghahramani and Jordan (1997). For overviews having a historical perspective see Ephraim and Merhav (2002) and Kouemou (2011).

On the other hand, the first relevant piece of literature which is specifically devoted to LM models for longitudinal data is represented by the book of Wiggins (1973), even if he had already presented some basic ideas in his PhD thesis that dates back to the 1950s (see Wiggins, 1955). Wiggins presented different LM models as extensions of the MC model, which take into account that the observed change occurring between two time occasions is typically a mixture of a true change and a spurious change due to measurement error.

The initial LM formulation of Wiggins (1973) has been developed in several directions and in connection with applications in psychology, sociology, and medicine. The first advance concerns the use of covariates. As for other latent variable models, the covariates may be included either in the measurement model or in the latent model. The choice depends on the type of application. The first approach was developed in Bartolucci and Farcomeni (2009), who proposed a multivariate LM model in which the conditional distribution of the response variables given the latent process is reparametrized through a multivariate logistic transformation (McCullagh and Nelder, 1989; Glonek and McCullagh, 1995). This link function is based on marginal logits, log-odds ratios, and similar higher-order effects. The second approach was adopted by Vermunt et al. (1999), who proposed to model initial and transition probabilities of the latent

Markov process through a series of multinomial logit regression models depending on time-constant and time-varying covariates. More recently, Bartolucci et al. (2007) extended this approach to the case of more than one response variable. Bartolucci and Pennoni (2007) also allowed transition probabilities to depend on lagged response variables.

Other interesting extensions are to multilevel data, where sample units are collected into clusters. A method based on fixed effects to represent the factors common to all units in the same cluster was proposed by Bartolucci et al. (2009). A formulation based on random parameters having a discrete distribution was instead proposed by Bartolucci et al. (2011). This extension is related to the mixed LM model (van de Pol and Langeheine, 1990) and to the LM model with random effects (Altman, 2007). These formulations for multilevel data are related to extended LM models in which the parameters are allowed to vary in different latent subpopulations. This approach is the basis of *latent transition analysis* (Bye and Schechter, 1986; Langeheine, 1988; Collins and Wugalter, 1992; Kaplan, 2008).

Finally, we have to mention that LM models have been successfully applied in several fields. We mention, in particular, the following fields of application:

- *psychological and educational measurement*: see Vermunt et al. (1999) for an application based on a German educational panel study among secondary school pupils repeatedly interviewed about school grades and their interest in certain subjects and Bartolucci and Solis-Trapala (2010) for an application based on data coming from a developmental study about certain psychological attitudes in early childhood; see Humphreys and Janson (2000) for a related study;

- *medicine and health*: see Auranen et al. (2000) for a study about transmission of pneumococcal carriage based on data collected in a sample of families with young children, Cook et al. (2000) for an application about the joint classification distribution of multiple diagnostic tests applied repeatedly over time based on a lung study, Bartolucci et al. (2009) for a study about the evolution of psycho-physic conditions of a sample of elderly individuals hosted in certain Italian nursing homes, and Rijmen et al. (2008) for a study about the course of emotions among anorectic patients;

- *criminology*: see Bijleveld and Mooijaart (2003) for an analysis of empirical recidivism data of juvenile delinquents from the Netherlands and Bartolucci et al. (2007) for a related study based on conviction histories of a cohort of offenders who were born in

England and Wales (these data are also considered in this book, see Section 1.4.2);

- *marketing and related fields*: see Poulsen (1990) for an application about brand choice behavior of a group of customers and Paas et al. (2007) for a study about ownership of financial products of households in the Netherlands;
- *labor market*: see Bartolucci and Farcomeni (2009) for a study about fertility and female participation in the labor market based on dataset coming from PSID (also considered in this book, see Section 1.4.3) and Richardson et al. (2011) for a study about the relation between labor market activity and health status based on longitudinal data from New Zeland.

1.3 Alternative approaches

In the statistical literature, several models have been developed which have a structure similar to that of LM models or have a similar range of applications. Interesting overviews in this regard are provided by Muthén (2004) and Vermunt (2010). In particular, two important classes of alternative models are based on the *latent growth approach* and on the *mixture Markov model*.

The latent growth approach is based on formulating an equation that, for each sample unit, relates the distribution of the response variables to covariates associated to the time occasion and is based on individual-specific parameters. Moreover, these individual-specific parameters are assumed to have a Gaussian distribution with mean that may also depend on available covariates. Models formulated on the basis of this approach have different names, such as *latent growth models* or *latent curve models* (Nagin, 1999). The basic formulation, based on continuous random effects affecting the distribution of every response variable, has been extended by using random effects that follow a mixture of Gaussian random variables or have a discrete distribution, leading to the so-called *growth mixture models* and *latent class growth models* (Muthén and Shedden, 1999). For a review of this approach see Bollen and Curran (2006).

The mixture Markov approach is based on separate MC models for latent subpopulations. Each subpopulation has specific initial and transition probabilities of the Markov chain, and the probability of belonging to a certain subpopulation typically depends on individual covariates. This approach was developed by Dias and Vermunt (2007) for an application in market segmentation. A particular case is the mover-stayer model of Goodman (1961).

Finally, we have to mention that in the econometric literature other latent variable models have been proposed which are alternative to LM models and have a similar structure. In particular we consider models for longitudinal data with individual covariates, in which the unobserved heterogeneity is represented by a latent autoregressive first-order process; see Heiss (2008). This model may be seen as the continuous counterpart of an LM model with individual covariates in the measurement model. In fact, the two models share the same basic assumptions, but in the first the latent variables are continuous, whereas in the second they are discrete.

1.4 Example datasets

In order to illustrate the models and the approaches introduced in this book, we will rely on a series of applications based on datasets which are available in the literature and were collected by longitudinal surveys. They cover social and economic fields. These datasets are described in the following subsections. For each dataset, we report the main descriptive statistics, and we indicate which relevant research questions may be addressed by its analysis through an LM model.

1.4.1 Marijuana consumption dataset

This dataset is based on 5 annual waves of the National Youth Survey (Elliot et al., 1989) and concerns 237 individuals who were aged 13 years in 1976. The *use of marijuana* was measured by an ordinal response variable for each wave, having the following three categories:

- "never in the past year" (coded as 0);
- "no more than once in a month in the past year" (coded as 1);
- "more than once a month in the past year" (coded as 2).

The distribution of the response variable for each wave is reported in Table 1.1. These data have been used for empirical demonstrations by Lang et al. (1999), Vermunt and Hagenaars (2004), and Bartolucci (2006). Similar data based on the Youth Risk Behavior Survey of 2005 and on the National Longitudinal Study of Adolescence Health have been used by Collins and Lanza (2010) for illustrative examples.

TABLE 1.1
Frequency distribution of the response variable for each wave

	Response		
Wave	0	1	2
1	218	14	5
2	195	27	15
3	167	41	29
4	156	41	40
5	138	52	47

The substantive research question, which can be addressed by the analysis of this dataset, concerns the evolution of the marijuana use with age and the strength of the dependence of the consumption in a given year on the consumption in the previous period.

The above aspects may be studied by a suitably formulated LM model that, as mentioned above, also accounts for measurement errors, whose existence is expected in a survey about drug consumption.

1.4.2 Criminal conviction history dataset

This dataset concerns the conviction histories of a cohort of 11,400 offenders (9,232 males and 2,168 females) who were born in England and Wales in 1953. The offenders were followed from the age of criminal responsibility, 10 years, until the end of 1993.

In particular, we use the data as organized by Bartolucci et al. (2007). Then, we consider 6 age bands: 10–15, 16–20, 21–25, 26–30, 31–35, and 36–40 years. For every age band, 10 binary response variables are available, which refer to specific offense groups. Each response variable is equal to 1 if the subject has been convicted for a crime of the corresponding group and to 0 otherwise. The considered offense groups are described in the book Research Development and Statistics Directorate (1998) and are the following:

1. *violence against the person*,
2. *sexual offenses*,
3. *burglary*,
4. *robbery*,
5. *theft and handling stolen goods*,
6. *fraud and forgery*,

7. *criminal damage*,
8. *drug offenses*,
9. *motoring offenses*,
10. *other offenses*.

The distribution of the response variables is reported, for each age band, in Table 1.2. It is important to note that the condition for the inclusion in the dataset is to be convicted at least once for one of the crimes during the period of observation. Therefore, a proper analysis of these data must take into account the overall number of reference subjects, convicted at least once and never convicted. According to Prime et al. (2001), a reliable estimate for this population size is 684000 (352000 males and 332000 females). Note that the data that are available to us are referred to one criminal out of every 13.

TABLE 1.2
Frequency of crimes for each offense group and age band separated by gender

	Offense group (males)									
Age band	1	2	3	4	5	6	7	8	9	10
1	99	48	1035	24	1941	31	378	12	0	243
2	774	138	1128	84	3027	295	756	311	28	1145
3	770	138	640	60	1932	389	612	403	172	527
4	572	97	375	41	1209	353	399	269	44	273
5	357	69	194	27	774	231	256	172	40	170
6	230	55	75	17	385	161	137	101	12	152
	Offense group (females)									
Age band	1	2	3	4	5	6	7	8	9	10
1	11	0	37	2	324	5	9	1	0	14
2	52	2	50	4	531	87	23	39	0	72
3	67	8	24	2	411	100	33	44	1	37
4	53	1	9	0	363	101	38	33	1	27
5	34	1	8	3	214	68	29	23	0	21
6	35	3	1	3	108	44	22	15	0	15
	Offense group (overall)									
Age band	1	2	3	4	5	6	7	8	9	10
1	110	48	1072	26	2265	36	387	13	0	257
2	826	140	1178	88	3558	382	779	350	28	1217
3	837	146	664	62	2343	489	645	447	173	564
4	625	98	384	41	1572	454	437	302	45	300
5	391	70	202	30	988	299	285	195	40	191
6	265	58	76	20	493	205	159	116	12	167

Scientific questions that may be addressed by the analysis of these data concern the detection of different groups of subjects having different patterns of criminal behavior and how this criminal behavior evolves across time, even depending on gender. A suitably formulated LM model may then represent a valid tool for the analysis of such data and to address these questions. An approach of analysis of this type was initially proposed by Bijleveld and Mooijaart (2003). In particular, they applied an LM model to data having a structure similar to the above one, but in which the criminal behavior is represented by only one response variable for each time period. Bartolucci et al. (2007), instead, applied a multivariate LM model with covariates, which directly uses the binary response variables corresponding to each type of crime and allows for more sophisticated analyses.

It is worth noting that the criminology literature has extensively focused on identifying groups of individuals with similar patterns of crimes and how a subject may evolve in his/her behavior or, in other terms, move between these groups depending on his/her background characteristics; see among others Nagin and Land (1993) and D'Unger et al. (1998). One of the key contributions in this field is due to Moffitt (1993), who proposed a distinction between the trajectories of "life-course persistents" versus "adolescence limiteds." Other relevant references are Soothill et al. (2002) and Francis et al. (2004).

1.4.3 Labor market dataset

This dataset was extracted from the database developed within the PSID, which is primarily sponsored by the National Science Foundation, the National Institute of Aging, and the National Institute of Child Health and Human Development. The study is conducted by the University of Michigan.[4]

The data used in our applications concern $n = 1,446$ women who were followed from 1987 to 1993. There are two binary response variables and several covariates observed for each year. The response variables are

- *fertility*: indicating whether a woman had given birth to a child in the year;
- *employment*: indicating whether the woman was employed during the year.

[4] See the website `http://psidonline.isr.umich.edu/`

The covariates are

- *race*: dummy variable equal to 1 for a black woman;
- *age*: in 1986;
- *education*: years of schooling;
- *child 1–2*: number of children in the family aged between 1 and 2 years (referred to the previous year);
- *child 3–5*: as above for children aged between 3 and 5 years;
- *child 6–13*: as above for children aged between 6 and 13 years;
- *child 14–*: as above for children aged at least 14 years;
- *income of the husband*: in dollars (referred to the previous year).

In Table 1.3 we report some descriptive statistics of the distributions of the available covariates, whereas in Table 1.4 we report the joint frequencies of the two response variables for every year of observation.

The main scientific question concerns the association between fertility and employment, taking also into account the unobserved heterogeneity between subjects. As shown by Bartolucci and Farcomeni (2009), a multivariate LM model with covariates affecting the conditional response probabilities, given the latent state, allows for this analysis. In particular, through a model of this type we can take into account that the effect of unobserved covariates may be time varying and disentangle the *contemporary dependence* between the two response variables from

TABLE 1.3
Descriptive statistics for the distributions of the covariates

Covariate	%	Mean	St.dev.
Race	23.86		
Age		29.55	4.61
Education		13.14	2.06
Child 1-2	57.05		
Child 3-5	66.39		
Child 6-13	75.10		
Child 14-	39.83		
Income of the husband		30.40	25.07

Note: For Child 1–2 we report the frequency of subjects for whom this variable is equal to 1 for at least one year in the period of observation; similar frequencies are reported for Child 3–5, Child 6–13, and Child 14–.

TABLE 1.4
Frequency of every response configuration (fertility, employment) for each year of observation

Year	Response configuration			
	(0,0)	(1,0)	(0,1)	(1,1)
1987	404	70	884	88
1988	425	55	902	64
1989	400	63	919	64
1990	377	45	969	55
1991	400	40	969	37
1992	424	23	965	34
1993	418	28	985	15

the *sequential dependence*. These issues are of great interest in labor economics; see Hyslop (1999) and Carrasco (2001) and the references therein.

1.4.4 Student math achievement dataset

This dataset derives from the administration of test items, aimed at assessing proficiency in mathematics, to students attending public and nonpublic middle schools in an Italian region. The data have been collected by the Regional Research Institute on Education within a regional project developed as a pilot study for the Italian Institute for the Evaluation of the Education System.

The items were administered at the end of each of the three years of school to 1,246 students, from 13 public and 7 nonpublic middle schools, who progressed from Grade 6 to Grade 8. Moreover, the number of items administered to these students, all dichotomously scored, were 28 at the end of the first year (Grade 6), 30 at the end of the second year (Grade 7), and 39 at the end of the third year (Grade 8). Each questionnaire included some items out of the Grade level for vertical scaling.

The dataset also includes the following covariates at student's level:

- *Father's educational level*: with four categories ("primary school," "middle school," "high school," and "college degree");

- *Mother's educational level*: with four categories defined as above.

Moreover, the following covariates at school level are included

- *Type of school*: "public" or "nonpublic,"
- *Students/teachers ratio*,
- *Years since school opened*.

The percentage distribution of the student-level covariates is reported in Table 1.5, whereas the number of correct responses for each item is reported in Table 1.6.

TABLE 1.5
Descriptive statistics for the student-level covariates

Variable	%
Father education	
primary school	3.4
middle school	22.6
high school	38.3
college degree	27.0
missing response	8.7
Mother education	
primary school	3.1
middle school	2.6
high school	43.4
college degree	23.3
missing response	7.8

The main issues in analyzing these data concern the evolution of the ability in mathematics of the pupils during the three years of the study, taking in particular into account how this evolution is affected by the individual covariates and by the type of school. For this aim, Bartolucci et al. (2011) formulated a multilevel LM model for multivariate longitudinal data, which is based on a Rasch parametrization (Rasch, 1961) for the conditional distribution of the probabilities of success given the latent state. In this way, each latent state is associated with a different level of ability.

The above approach represents an alternative to the approach based on *value added models*, which was first proposed by Bryk and Weisberg (1976); see also Goldstein et al. (2007) for recent developments. However, in our view the LM approach described above allows for more sophisticated analyses mainly because, in a single framework, we have a sensible measurement model, which relates the response success probabilities to the ability level, and a latent model which includes individual covariates and random effects to take into account the multilevel structure of the data. Through these random effects we can also evaluate the school performance, even depending on covariates such as *type of school*.

TABLE 1.6
Number of correct responses for every item administered at the end of each year of middle school

	Number of correct responses		
Item	grade 6	grade 7	grade 8
1	1035	806	887
2	332	1141	806
3	951	752	864
4	1133	785	563
5	674	950	786
6	496	755	848
7	684	1016	762
8	755	813	255
9	695	985	611
10	748	510	714
11	485	574	297
12	634	940	690
13	705	380	754
14	270	485	451
15	543	1083	742
16	614	629	830
17	901	652	480
18	353	207	934
19	618	373	551
20	906	453	550
21	493	392	849
22	903	490	478
23	528	477	909
24	579	924	507
25	568	804	587
26	424	639	879
27	796	672	1091
28	508	886	812
29		643	838
30		512	925
31			827
32			779
33			494
34			686
35			435
36			856
37			872
38			784
39			496

2

Background on latent variable and Markov chain models

2.1 Introduction

Latent Markov (LM) models are latent variable models tailored to the analysis of longitudinal data, typically used when the response variables are categorical. These models make use of time-specific latent variables, which are assumed to be discrete, and then represent a rather sophisticated class of latent variable models. In fact, even in its simplest formulation, an LM model may be seen as a generalization of the latent class (LC) model, which is a well-known model for classifying a sample of subjects on the basis of a set of categorical responses. Another possible interpretation of an LM model is as an extension of the Markov chain (MC) model allowing for measurement errors.

The aim of this chapter is to provide the reader with basic concepts about latent variable models and, in particular, about the LC model and the MC model. These models represent a useful paradigm to explain several basic concepts.

We only discuss those concepts that are useful for illustrating LM models, and we keep the presentation as simple as possible. For more general overviews on latent variable models, suggested readings are Skrondal and Rabe-Hesketh (2004) and Bartholomew et al. (2011) whereas about the MC model, suggested readings are Taylor and Karlin (1998), and Frees (2004).

2.2 Latent variable models

As stated by Skrondal and Rabe-Hesketh (2004) “latent variables pervade modern mainstream statistics and are widely used in different disciplines such as medicine, economics, engineering, psychology, geography, marketing and biology" (p. 1). The techniques involving latent variables have risen and grown almost exclusively within the framework of the social and behavioral sciences, but applications in other fields are becoming more and more common. These techniques have different names in a wide literature that is extended over almost a century.

Given the wide range of applications, it is difficult to formulate a simple and general definition for latent variable models. We propose a definition which is valid in our context. *A latent variable model is a model which relies on specific assumptions on the conditional distribution of the response variables, given one or more variables which are not directly observable (latent variables).* These models typically assume a simplified dependence structure for the response variables given the latent variables. In this regard, a fundamental assumption is that of *local independence*, according to which the response variables are conditionally independent given the latent variables. The motivation behind this assumption is that the latent variables represent the only explanatory factor of the outcomes, since the latter ones provide a measure of the first ones.

Different ways to classify latent variable models are available in the literature, depending on the number of latent variables and the nature of these variables and the response variables. These classifications essentially distinguish between discrete and continuous variables. In any case, it is usually possible to disentangle two components of a latent variable model, which are formulated through specific assumptions:

1. *measurement model*: it describes the conditional distribution of the response variables given the latent variables;

2. *latent model*: it describes the distribution of the latent variables.

By jointly considering the two above components and after some simple rules, we obtain the so-called *manifest distribution*, that is, the marginal distribution of the response variables, once the latent variables have been integrated out.

As already mentioned, we are here interested in models based on latent variables having a discrete distribution. These are referred to as *discrete latent variable models* and have many points in common with *finite-mixture models* (McLachlan and Peel, 2000). In this case, the different configurations of the latent variables identify different subpopulations of subjects or, more generally, of statistical units. These subpopulations are usually referred to as *latent classes.* These models find natural application in the presence of categorical response variables.

With reference to a random unit drawn from the population of interest, let $Y_1, \dots, Y_r$ denote the response variables, which are collected in the random vector $\boldsymbol{Y}$. These are random variables and are denoted by capital letters. In addition, realizations of random variables and random vectors will be denoted by small letters, and this convention will be used throughout the book. Note that $Y_1, \ldots, Y_r$ may correspond to variables having a different nature observed at the same occasion, in the case of cross-sectional data, or to repeated observations at different time occasions, so as to include the case of longitudinal data. Moreover, let $U_1, \ldots, U_l$ denote the corresponding latent variables, which are collected in the vector $\boldsymbol{U}$. Typically, l (number of latent variables) is much smaller than r (number of response variables); in many applications we have $l = 1$. Then, the measurement model concerns the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{U}$ and the corresponding probability mass function (or density function in the continuous case) is denoted by

$$f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) = f(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{U} = \boldsymbol{u}),$$

where $\boldsymbol{y}$ denotes a realization of $\boldsymbol{Y}$ and $\boldsymbol{u}$ denotes a realization of $\boldsymbol{U}$. This distribution depends on specific parameters that, for the moment, are not explicitly indicated. The notation adopted above for the probability mass (or density) function will be adopted throughout the book. Moreover, the latent model corresponds to the (*a priori*) distribution of the latent vector $\boldsymbol{U}$ and formulates assumptions on the probability mass (or density) function

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = f(\boldsymbol{U} = \boldsymbol{u}).$$

This distribution also depends on specific parameters, which are not here explicitly indicated.

As already mentioned, we mainly consider models based on discrete latent variables. For these models, the manifest distribution of $\boldsymbol{Y}$ has probability mass (or density) function which depends on $f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u})$ and $f_{\boldsymbol{U}}(\boldsymbol{u})$ as follows

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{\boldsymbol{u}} f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) f_{\boldsymbol{U}}(\boldsymbol{u}), \tag{2.1}$$

where the sum $\sum_{\boldsymbol{u}}$ is over all possible configurations of $\boldsymbol{U}$. This expression clarifies that, at least in the *unidimensional case*, where only one latent variable is assumed to exist ($l = 1$), a discrete latent variable model corresponds to a finite mixture model (Titterington et al., 1985; McLachlan and Peel, 2000; Lindsay, 1995). Note that, under the assumption of local independence, we have

$$f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) = \prod_{j=1}^{r} f_{Y_j}(y_j|\boldsymbol{u}), \tag{2.2}$$

where y_j is the j-th element of $\boldsymbol{y}$ and $f_{Y_j|\boldsymbol{U}}(y|\boldsymbol{u})$ is the probability mass (or density) function of the conditional distribution of the single response variable Y_j given $\boldsymbol{U}$, with y denoting a possible realization of this variable. This assumption then leads to a strong simplification of the model and has an easy interpretation that will be clarified in the following.

Another fundamental concept in this literature is that of *a posteriori* distribution of the latent variables, that is, the conditional distribution of the latent variables given a certain response configuration. From the Bayes' Theorem, the probability mass function of this distribution is given by

$$f_{\boldsymbol{U}|\boldsymbol{Y}}(\boldsymbol{u}|\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u})f_{\boldsymbol{U}}(\boldsymbol{u})}{f_{\boldsymbol{Y}}(\boldsymbol{y})}. \tag{2.3}$$

Once a latent variable model has been estimated, these probabilities are used to assign each subject to a certain latent variable configuration or, equivalently, latent class.

As an example, consider the case in which the response variables correspond to different items and indicators reflecting the quality-of-life of an elderly subject. The items concern the activity of daily living, whereas the indicators are related to certain clinical measures. In the present framework, the dependence structure between these variables is simplified by introducing one or a few latent variables. The different configurations of latent variables are interpreted as different levels of the quality-of-life, an individual characteristic which is only indirectly observable through the responses to these items and indicators. This interpretation is enforced by the assumption of local independence. In fact, this assumption states that if we knew the latent variable configuration of a subject, the response given to an item would not help to predict the response that the subject may provide to another item. All the information necessary to predict the responses is contained in the latent variables, which then correspond to the "true" quality-of-life level. It has to be stressed that this fundamental assumption does not state that the item responses are (marginally) independent, but that they are only conditionally independent given the latent variables.

It is worth noting that a latent variable model may also include individual covariates. These covariates may be included in the measurement model, so that they affect the conditional distribution of the response variables given the latent variables, or in the latent model, so that they affect the distribution of the latent variables. In the first case, the latent variables have the role of accounting for the *unobserved heterogeneity* between subjects, that is, the heterogeneity that cannot be explained by the observable covariates. Therefore, this formulation is adopted when the interest is in estimating the direct effect of these covariates on the response variables. In the second case, instead, the main interest is in understanding how the covariates affect the unobservable characteristic that is measured by the response variables.

Usually, covariates are not assumed to affect both the conditional distribution of the response variables given the latent variables and the distribution of the latent variables, because of the difficulties in interpreting and estimating the resulting model. In the following, we provide more details on this point with reference to the LC model.

Let $\boldsymbol{X}$ denote the (column) vector of individual covariates. If the covariates are included only in the measurement model, we denote by $f_{\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{x})$ the corresponding probability mass (or density) function that, under the assumption of conditional independence, may be factorized as in (2.2) on the basis of the functions $f_{Y_j|\boldsymbol{U},\boldsymbol{X}}(y|\boldsymbol{u},\boldsymbol{x})$. When the covariates are only included in the latent model, we denote by $f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x})$ the corresponding probability mass (or density) function. In any case, by manifest distribution of the response variables, we mean the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$. With discrete latent variables, this distribution has probability mass (or density) function

$$f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{u}} f_{\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{x}) f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x}), \tag{2.4}$$

where, depending on the adopted assumptions, we may have that $f_{\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{x}) = f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u})$ or $f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x}) = f_{\boldsymbol{U}}(\boldsymbol{u})$. Similarly, the posterior distribution of $\boldsymbol{U}$ is conditional on both $\boldsymbol{X}$ and $\boldsymbol{Y}$, and corresponds to

$$f_{\boldsymbol{U}|\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{u}|\boldsymbol{x},\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{x}) f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x})}{f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})}. \tag{2.5}$$

In the following, before illustrating the LC model, which is a particular latent variable model, we illustrate the Expectation-Maximization (EM) algorithm, which is typically used for maximum likelihood estimation of latent variable models.

2.3 Expectation-Maximization algorithm

In this section, we briefly review the EM algorithm, which is undoubtedly the main tool to estimate a latent variable model, especially when based on discrete latent variables. This algorithm was derived by Baum and colleagues in a series of papers specifically for hidden Markov models (Baum and Petrie, 1966; Baum and Egon, 1967; Baum et al., 1970), and then it was put in a more general context in the widely cited paper of Dempster et al. (1977). Exhaustive overviews on this algorithm and its extensions are presented by Watanabe and Yamaguchi (2004) and McLachlan and Krishnan (2008).

What we obtain from the EM algorithm is the maximum likelihood estimate of the parameters of the model on the basis of an observed sample drawn from the population of interest. In this regard, it is convenient to disentangle the case of absence of individual covariates, in which for all subjects we assume the same distribution of the response variables, from the case in which these covariates are present.

In absence of individual covariates, let $\boldsymbol{y}_i$ denote the observed response configuration for subject i in a sample of n independent units, so that $i = 1, \ldots, n$. Then, the model log-likelihood is equal to

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{Y}}(\boldsymbol{y}_i), \tag{2.6}$$

where $\boldsymbol{\theta}$ is the vector of model parameters and $f_{\boldsymbol{Y}}(\boldsymbol{y})$ is the manifest probability (or density) of $\boldsymbol{Y}$, already defined in (2.1). An equivalent and more convenient representation is

$$\ell(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}} n_{\boldsymbol{y}} \log f_{\boldsymbol{Y}}(\boldsymbol{y}), \tag{2.7}$$

where the sum $\sum_{\boldsymbol{y}}$ is over all response configurations observed at least once and $n_{\boldsymbol{y}}$ is the frequency of configuration $\boldsymbol{y}$ in the sample, that is,

$$n_{\boldsymbol{y}} = \sum_{i=1}^{n} I(\boldsymbol{y}_i = \boldsymbol{y}).$$

In the above expression, $I(\cdot)$ is the indicator function equal to 1 if its argument is true and to 0 otherwise. Expression (2.7) for the log-likelihood is more convenient to use than expression (2.6) because the number of distinct response configurations that are observed is always smaller than or equal to the sample size n. This implies certain advantages in performing the steps of the EM algorithm. Then, we will adopt this formulation in the following.

From the perspective of the EM algorithm, the problem of estimating a latent variable model is cast into the problem of estimating a statistical model in the presence of missing data. In our context, the *missing data* correspond to the vector of latent variables $\boldsymbol{u}_i$ for every subject i, and then the so-called *complete data* correspond to the pairs $(\boldsymbol{u}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$. Therefore, the *complete data log-likelihood* is given by

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{u}_i, \boldsymbol{y}_i),$$

where $f_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{u},\boldsymbol{y}) = f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u})f_{\boldsymbol{U}}(\boldsymbol{u})$ refers to the joint distribution of $\boldsymbol{U}$ and $\boldsymbol{Y}$. Then, we have

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}_i|\boldsymbol{u}_i) + \sum_{i=1}^{n} \log f_{\boldsymbol{U}}(\boldsymbol{u}_i).$$

Note that the incomplete (or observed) data correspond to the vectors $\boldsymbol{y}_i$, $i = 1, \ldots, n$, and therefore $\ell(\boldsymbol{\theta})$ is usually referred to as *incomplete data log-likelihood*.

The alternative expression for the complete data log-likelihood, which is used in practice, is

$$\ell^*(\boldsymbol{\theta}) = \sum_{\boldsymbol{u}}\sum_{\boldsymbol{y}} a_{\boldsymbol{u}\boldsymbol{y}} \log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) + \sum_{\boldsymbol{u}} b_{\boldsymbol{u}} \log f_{\boldsymbol{U}}(\boldsymbol{u}), \qquad (2.8)$$

where $b_{\boldsymbol{u}}$ is the number of sample units having latent variable configuration $\boldsymbol{u}$ and $a_{\boldsymbol{u}\boldsymbol{y}}$ is the number of these sample units also having response configuration $\boldsymbol{y}$, that is,

$$a_{\boldsymbol{u}\boldsymbol{y}} = \sum_{i=1}^{n} I(\boldsymbol{u}_i = \boldsymbol{u}, \boldsymbol{y}_i = \boldsymbol{y}), \quad b_{\boldsymbol{u}} = \sum_{i=1}^{n} I(\boldsymbol{u}_i = \boldsymbol{u}).$$

Since the latent configuration of each subject is not known, the EM algorithm maximizes $\ell(\boldsymbol{\theta})$ by alternating the following two steps until convergence:

- **E-step:** compute the conditional expected values of $\ell^*(\boldsymbol{\theta})$ given the observed data and the current value of the parameters;
- **M-step:** update $\boldsymbol{\theta}$ with the parameter vector which maximizes the above expected value.

Both steps are usually simple to implement. In particular, the E-step is typically based on the posterior distribution of the latent variables. In fact, this expected value is obtained by substituting, in expression (2.8), each unknown frequency $a_{\boldsymbol{u}\boldsymbol{y}}$ and $b_{\boldsymbol{u}}$ with the corresponding conditional expected value computed on the basis of the posterior probabilities in (2.3). The M-step is usually based on a series of maximizations which involve different blocks of parameters; these maximizations are based on explicit rules or on simple iterative algorithms of Newton-Raphson (NR) type.

In the presence of individual covariates, for every subject i we observe the vector $\boldsymbol{x}_i$ and the vector $\boldsymbol{y}_i$. Then, the model log-likelihood has the expression

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}_i|\boldsymbol{x}_i) = \sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}} n_{\boldsymbol{x}\boldsymbol{y}} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}), \qquad (2.9)$$

where the sum $\sum_{\boldsymbol{x}}$ is over all covariate configurations observed at least once and

$$n_{\boldsymbol{x}\boldsymbol{y}} = \sum_{i=1}^{n} I(\boldsymbol{x}_i = \boldsymbol{x}, \boldsymbol{y}_i = \boldsymbol{y})$$

is the joint frequency of the covariate configuration $\boldsymbol{x}$ and response configuration $\boldsymbol{y}$. The corresponding expression for the complete data log-likelihood is

$$\begin{aligned} \ell^*(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log f_{\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{X}}(\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{x}_i) + \sum_{i=1}^{n} \log f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}_i|\boldsymbol{x}_i) \\ &= \sum_{\boldsymbol{u}} \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}} a_{\boldsymbol{u}\boldsymbol{x}\boldsymbol{y}} \log f_{\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{x}) \\ &+ \sum_{\boldsymbol{u}} \sum_{\boldsymbol{x}} b_{\boldsymbol{u}\boldsymbol{x}} \log f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x}), \end{aligned}$$

which is based on the frequencies

$$a_{\boldsymbol{u}\boldsymbol{x}\boldsymbol{y}} = \sum_{i=1}^{n} I(\boldsymbol{u}_i = \boldsymbol{u}, \boldsymbol{x}_i = \boldsymbol{x}, \boldsymbol{y}_i = \boldsymbol{y}), \quad b_{\boldsymbol{u}\boldsymbol{x}} = \sum_{i=1}^{n} I(\boldsymbol{u}_i = \boldsymbol{u}, \boldsymbol{x}_i = \boldsymbol{x}).$$

Even with individual covariates, the structure of the EM algorithm is the same as above and is based on alternating the E-step and the M-step until convergence.

Before going into a detailed illustration of the EM algorithm for the LC model, we have to recall that the main advantage of this algorithm with respect to direct maximization algorithms of the likelihood function, such as those of NR type, is the simplicity of implementation. In fact, NR type algorithms require to compute at least the first derivative vector of $\ell(\boldsymbol{\theta})$ and this may not be easy. Moreover, these algorithms are known to have difficulties in converging when applied to discrete latent variable models because of the log-likelihood function complexity.

On the other hand, the EM algorithm has two important drawbacks. The first is the slowness to converge. Typically, we observe that the first steps of the algorithm considerably increase the likelihood, but many iterations are needed for improving a solution close to the maximum likelihood estimate. Another problem in using the EM algorithm is related to the fact that, differently from other algorithms such as the NR or the Fisher-scoring, it uses neither the observed nor the expected information matrix. We recall that these matrices are suitable transformations of the second derivative matrix of the (incomplete data) log-likelihood. From these matrices we obtain standard errors for the parameter estimates, as we discuss in the following section.

Finally, we have to recall that the likelihood of a discrete latent variable model is typically multimodal, since it may have more local maxima. Obviously, in the presence of more local maxima, the EM algorithm converges to one of them that is not ensured to be the global maximum. In order to increase the chance that the point at convergence is the global maximum, we have to properly initialize the algorithm. Usually, a multiple-try strategy is adopted which is based on combining a deterministic rule with one or more random rules. Typically, the first is a simple rule, which leads to a reasonable guess of the parameter values obtained by fitting a simplified version of the adopted model. Once a multiple-try strategy is adopted, as the maximum likelihood estimate of the parameters, we take the value corresponding to the highest log-likelihood at convergence of the EM algorithm.

2.4 Standard errors

Once parameter estimates have been computed, standard errors are commonly associated to these estimates. Each standard error may be seen as a measure of precision of the estimate to which it refers, and it is typically used to test hypotheses on the corresponding parameter and to obtain confidence intervals.

As already mentioned, the general method to obtain standard errors is by the inverse of the observed or the expected information matrix. In particular, the standard errors are computed as the square root of the elements in the main diagonal of one of these matrices. Another method is based on parametric or nonparametric bootstrap which consists of repeatedly drawing samples from the observed sample (nonparametric version) or from the estimated model (parametric version) and computing the maximum likelihood estimate for every bootstrap sample. Then the standard error corresponding to a parameter estimate is found through the standard deviation of the empirical distribution so obtained. Bootstrap methods have in general the advantage of providing more reliable values for the standard errors especially when the sample size is small or the parameter estimates are close to the boundary of the parameter space. On the other hand, bootstrap methods may be computationally intensive, since they require estimating the same model on several samples. For a complete overview see Efron and Tibshirani (1993), Davison and Hinkley (1997), and Chernick (2008).

Concerning latent variable models, both methods to obtain standard errors, from the information matrix and from bootstrap samples, have received adequate attention. In particular, regarding the information matrix there are methods to compute the observed information matrix from the EM algorithm output. In particular, Louis (1982) developed a procedure in which this matrix is expressed as the difference between two matrices corresponding to the "total information," which we would have if we knew the latent states, and the "missing information." However, the expression of the latter may be cumbersome to compute and other expressions are available; see Oakes (1999). See also Lystig and Hughes (2002) for an alternative method which is specific for HM models. Moreover, bootstrap methods for latent variable and finite mixture models have been studied by other authors, such as Feng and McCulloch (1996) and Zucchini and MacDonald (2009, Chapter 3) specifically for HM models.

We will provide more details about the above methods in the following chapters specifically with reference to the LM models, whereas we do not enter into details in this chapter since it provides only an introduction on latent variable and other models. Consequently, in the applications that we will illustrate in the following, we will report parameter estimates without standard errors.

2.5 Latent class model

The LC model is one of the most well-known latent variable models. This model is typically used to classify a sample of subjects on the basis of a series of categorical response variables. It assumes the existence of only one discrete latent variable with a number of levels which is usually chosen on the basis of the data. In this way, the model helps to gain a deeper understanding of the observed relationships between the response variables.

The LC model was initially developed by Lazarsfeld (1950) and Green (1951); see also Lazarsfeld and Henry (1968) and Haberman (1974). Issues related to the maximum likelihood estimation of this model were studied in detail by Goodman (1974). During the past two decades, great progress has been made about estimating and testing the LC model and its extensions; see Langeheine (1988) and Clogg (1995). Also see Goodman (2002) for an exhaustive review about the LC model and estimation methods for this model.

In the following, we illustrate the basic version of the LC model and more advanced versions of this model based on the use of suitable parametrizations, which also allow us to include individual covariates.

2.5.1 Basic version

As above, let $Y_1, \dots, Y_r$ denote the response variables, collected in the vector $\boldsymbol{Y}$, which are categorical with an arbitrary number of categories denoted by c. These categories are labeled from 0 to $c-1$, so as to easily include the case of binary responses when $c = 2$. What follows may be easily adapted to the case of response variables having a different number of categories, but we prefer to refer to the case of the same number of categories to simplify the notation.

In order to explain the dependence structure between these variables, the LC model assumes the existence of only one discrete latent variable U ($l = 1$). Being based on the assumption of local independence, the path diagram in Figure 2.1 results for the LC model.

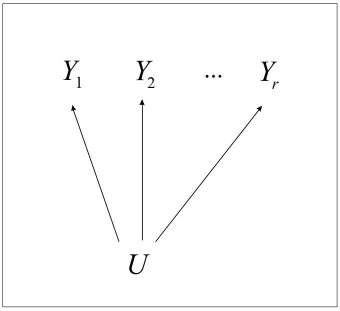

FIGURE 2.1
Path diagram for the LC model

The latent variable U has k levels, which are labeled from 1 to k. Each of these levels corresponds to a latent class in the population for which we have a specific weight, or *a priori probability*, and a specific conditional distribution of the response variables. More precisely, using the notation of Section 2.2, we define

$$\phi_{jy|u} = f_{Y_j|U}(y|u), \quad j = 1, \dots, r,\ u = 1, \dots, k,\ y = 0, \dots, c-1,$$

and

$$\pi_u = f_U(u), \quad u = 1, \dots, k,$$

where π_u is the weight of latent class u. The probabilities $\phi_{jy|u}$ and π_u are parameters to estimate, which, obviously, must be nonnegative and satisfy the constraint $\sum_{u=1}^{k} \pi_u = 1$ and the constraints $\sum_{y=0}^{c-1} \phi_{jy|u} = 1$, $j = 1, \dots, r$, $u = 1, \dots, k$. The number of free parameters is then

$$\#\text{par} = \underbrace{k-1}_{\pi_u} + \underbrace{kr(c-1)}_{\phi_{jy|u}},$$

that becomes

$$\#\text{par} = k - 1 + kr$$

in the case of binary response variables.

Using the above notation, we have

$$f_{\boldsymbol{Y}|U}(\boldsymbol{y}|u) = \prod_{j=1}^{r} \phi_{jy_j|u}$$

and the manifest distribution of $\boldsymbol{Y}$ becomes equal to

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{u=1}^{k} \left(\prod_{j=1}^{r} \phi_{jy_j|u} \right) \pi_u,$$

which directly derives from (2.1). Finally, the posterior probability that a subject with observed response configuration $\boldsymbol{y}$ belongs to latent class u is

$$f_{U|\boldsymbol{Y}}(u|\boldsymbol{y}) = \frac{\left(\prod_{j=1}^{r} \phi_{jy_j|u} \right) \pi_u}{f_{\boldsymbol{Y}}(\boldsymbol{y})}, \tag{2.10}$$

which can be derived from (2.3). This function is used to construct an allocation rule, according to which an individual is assigned to the class with the largest posterior probability.

2.5.2 Advanced versions

In certain applications, constraints may be posed on the parameters of the LC model presented above. The formulation of these constraints requires suitable parametrizations which may concern the measurement model or the latent model. Essentially the same parametrizations may be adopted in order to include in the model individual covariates, which are often available.

Constraints are typically assumed on the measurement model only. In the case of binary response variables, the most relevant constraint is that the conditional probabilities of success are equal to a certain value, such as 0, for a certain class. A more sophisticated example, which is illustrated in the following, is when these probabilities are parameterized as in the Rasch model (Rasch, 1961), which is the most well-known Item Response Theory (IRT) model (Hambleton and Swaminathan, 1985).

Example 1 — LC Rasch model. *Consider the parametrization*

$$\log \frac{\phi_{j1|u}}{\phi_{j0|u}} = \alpha_u - \psi_j, \quad j = 1, \ldots, r,\ u = 1, \ldots, k. \tag{2.11}$$

The resulting model is usually referred to as the LC Rasch model (Lindsay et al., 1991; Formann, 1995). This model is suitable for the analysis of data deriving from the administration of a set of dichotomously scored test items to a group of subjects, a situation that frequently arises in psychological and educational measurement. In such a case, α_u is interpreted as the ability of the subjects in the u-th latent class and ψ_j as the difficulty of the j-th item. Note that, in order to ensure model identifiability, we have to use one constraint, such as $\alpha_1 = 0$ or $\psi_1 = 0$.

We can also formulate multidimensional versions of the Rasch model, which are suitable in the case of items measuring different types of ability; see Holland and Rosenbaum (1986). In this case, a vector of abilities must be associated to each latent class; see Bartolucci (2007) for a detailed description.

It is worth noting that parametrization (2.11) implies the constraint of monotonicity because, under this parametrization, the latent classes can always be ordered so that

$$\phi_{j1|1} \leq \phi_{j1|2} \leq \cdots \leq \phi_{j1|k}, \quad j = 1, \ldots, r. \tag{2.12}$$

In this way, we can identify the latent classes according to increasing levels of the probability of success to all items. However, we have to consider that constraint (2.12) may be directly included into the LC model without the need of assuming parametrization (2.11). Therefore, a more flexible IRT model than the LC Rasch model results; see Bartolucci and Forcina (2005) for a detailed description.

An important development of the basic LC model is the incorporation of individual covariates, first proposed by Dayton and Macready (1988); see also Dayton and Macready (2002). This topic has also been developed by Bandeen-Roche et al. (1997) and Huang and Bandeen-Roche (2004).

We recall that we denote by $\boldsymbol{X}$ the vector of the covariates and by $\boldsymbol{x}$ a corresponding realization of this vector. In this setting, the distributions $f_{\boldsymbol{Y}|U,\boldsymbol{X}}(\boldsymbol{y}|u,\boldsymbol{x})$ and $f_{U|\boldsymbol{X}}(u|\boldsymbol{x})$ are modeled, through suitable parametrizations, as functions of the covariates and we use the notation

$$\begin{aligned}\phi_{jy|u\boldsymbol{x}} &= f_{Y_j|U,\boldsymbol{X}}(y|u,\boldsymbol{x}), \quad j = 1,\ldots,r,\ u = 1,\ldots,k,\ y = 0,\ldots,c-1,\\ \pi_{u|\boldsymbol{x}} &= f_{U|\boldsymbol{X}}(u|\boldsymbol{x}), \quad u = 1,\ldots,k.\end{aligned}$$

We have to choose between including these covariates into the measurement model or into the latent model. The choice essentially depends on the context of application; see also the discussion in Bartolucci and Forcina (2006). In any case, the manifest distribution of $\boldsymbol{Y}$ and the posterior distribution of U are now conditional on $\boldsymbol{X}$ and are obtained through the general formulae (2.4) and (2.5).

Two examples of covariates included in the measurement model are proposed below.

Example 2 — Logit model with discrete random effects. *In the case of binary response variables, consider the assumption*

$$\log\frac{\phi_{j1|u\boldsymbol{x}}}{\phi_{j0|u\boldsymbol{x}}} = \alpha_u + \psi_{1j} + \boldsymbol{x}'\boldsymbol{\psi}_{2j}, \tag{2.13}$$

which holds for $j = 1,\dots,r$ and $u = 1,\dots,k$, where ψ_{1j} and $\boldsymbol{\psi}_{2j}$ are response-specific parameters. Also assume that the a priori *probabilities of the latent classes do not depend on the covariates. Then, the parameters to estimate are α_u, ψ_{1j}, and $\boldsymbol{\psi}_{2j}$, together with the class weights π_u. One identifiability constraint, such as $\alpha_1 = 0$ or $\psi_{11} = 0$, is required.*

The model based on the above assumptions may be seen as a discrete version of a random-effects logit model and then it finds application when we want to take into account the unobserved heterogeneity between subjects. In applying this model, the main interest is in the parameters $\boldsymbol{\psi}_j$ which allow us to measure the direct effect of each covariate on the response variables; see McCulloch et al. (2008).

It is worth noting that we may extend the parametrization in (2.13) by using a specific vector of covariates $\boldsymbol{x}_j$ for each combination of subject and response variable. We can also include, among the covariates, the lagged response variables, relaxing in this way the assumption of local independence. It results in an LC version of the dynamic logit model (Hsiao, 2003) for longitudinal data. This model is typically applied to analyze labor market data and, in particular, to estimate the effect of the state dependence (Heckman, 1981). This means measuring the effect of having a job position in a given year on the probability of having the same job position in the following years, once observable covariates and subject-specific unobservable factors common to all time occasions are taken into account.

Example 3 — Proportional odds model with discrete random effects. *This is an extension of the above model for ordinal response variables with categories $0,\dots,c-1$, which is based on the assumption*

$$\log\frac{\phi_{jy|u\boldsymbol{x}} + \cdots + \phi_{j,c-1|u\boldsymbol{x}}}{\phi_{j0|u\boldsymbol{x}} + \cdots + \phi_{j,y-1|u\boldsymbol{x}}} = \alpha_{ju} + \psi_{1jy} + \boldsymbol{x}'\boldsymbol{\psi}_{2j}, \quad y = 1,\dots,c-1, \tag{2.14}$$

for $j = 1,\dots,r$ and $u = 1,\dots,k$. The resulting model may be seen as a discrete version of a proportional odds model (McCullagh, 1980) *with random effects having a discrete distribution. One identifiability constraint is also required in this case.*

About assumption (2.14), it is worth noting that the adopted parametrization is based on the so-called global logits, which are strongly related to cumulative logits (Agresti, 2002) and are formulated by comparing the following two probabilities

$$\begin{aligned} p(Y_j \geq y|U=u, \boldsymbol{X}=\boldsymbol{x}) &= \phi_{jy|u\boldsymbol{x}} + \cdots + \phi_{j,c-1|u\boldsymbol{x}}, \\ p(Y_j < y|U=u, \boldsymbol{X}=\boldsymbol{x}) &= \phi_{j0|u\boldsymbol{x}} + \cdots + \phi_{j,y-1|u\boldsymbol{x}}, \end{aligned}$$

for each cut-point $y = 1, \ldots, c-1$. Also note that the intercepts α_{ju} are specific to each cut-point, whereas, as in the formulation of McCullagh (1980), the parameters $\boldsymbol{\psi}_{2j}$ depend only on the response variable.

As an example of the alternative formulation, in which the covariates are included in the latent model and then affect the *a priori* distribution of the latent classes, consider the following.

Example 4 — Latent regression model. *Regardless of the number of response categories, assume the multinomial logit parametrization*

$$\log \frac{\pi_{u|\boldsymbol{x}}}{\pi_{1|\boldsymbol{x}}} = \gamma_{1u} + \boldsymbol{x}'\boldsymbol{\gamma}_{2u}, \quad u = 2, \ldots, k; \tag{2.15}$$

assume also that the conditional response probabilities do not depend on the covariates, that is, $\phi_{jy|u} = \phi_{jy|u\boldsymbol{x}}$ for all $\boldsymbol{x}$, so that the parameters γ_{1u} and $\boldsymbol{\gamma}_{2u}$ have to be estimated together with the probabilities $\phi_{jy|u}$.

The above formulation is of interest when we want to understand how the covariates affect the latent characteristic which is indirectly measured by the response variables; see Bandeen-Roche et al. (1997) and Huang and Bandeen-Roche (2004). For instance, the latent characteristic may correspond to the quality-of-life of an elderly subject which is affected by several factors, such as gender, age, and type of nursing home where the subject is hosted.

One may wonder about a formulation of the LC model which includes the individual covariates both in the measurement model, through a parametrization of type (2.13), and in the model for the latent variable, through a parametrization of type (2.15). In general, we advise against formulations of this type because the resulting model would be difficult to interpret and often nonidentifiable or almost nonidentifiable with consequent estimation problems. On the other hand, in many applications, models are of interest in which a constrained parametrization for the measurement (or latent) model is combined with a parametrization including individual covariates for the latent (or measurement) model. For instance, we can formulate an LC Rasch model, based on (2.11), in which the distribution of the latent variable depends on the covariates through parametrization (2.15).

What we have presented above are only some examples of the possible parametrizations for latent variable models, so as to allow the reader to understand the sense of more advanced formulations of the LC model. A general overview on the possible parametrizations will be given in Chapters 4 and 5, when similar extensions will be formulated with reference to the LM model, which is the main subject of this book.

2.5.3 Maximum likelihood estimation

In absence of individual covariates, let $\boldsymbol{y}_i$ denote the vector of responses observed for subject i, $i = 1, \ldots, n$, with elements $y_{i1}, \ldots, y_{ir}$. The log-likelihood $\ell(\boldsymbol{\theta})$ has the same expression as in (2.6) and (2.7).

In order to maximize $\ell(\boldsymbol{\theta})$, we can use the EM algorithm, which is based on the complete data log-likelihood. After some simple algebra, this reduces to

$$\ell^*(\boldsymbol{\theta}) = \sum_{j=1}^{r} \sum_{u=1}^{k} \sum_{y=0}^{c-1} a_{juy} \log \phi_{jy|u} + \sum_{u=1}^{k} b_u \log \pi_u,$$

where a_{juy} is the frequency of subjects that are in latent class u and respond by y at the j-th response variable, that is,

$$a_{juy} = \sum_{i=1}^{n} I(u_i = u, y_{ij} = y).$$

The expression for the complete data log-likelihood directly derives from (2.8) given the assumption of local independence.

In this setting, the E-step consists of computing the conditional expected value of each frequency a_{juy} and b_u. These expected values, denoted by $\hat{a}_{juy}$ and $\hat{b}_u$, are computed at the current value of the parameters through (2.10). In particular, we have

$$\begin{aligned} \hat{a}_{juy} &= \sum_{i=1}^{n} I(y_{ij} = y) f_{U|\boldsymbol{Y}}(u|\boldsymbol{y}_i). \\ \hat{b}_u &= \sum_{i=1}^{n} f_{U|\boldsymbol{Y}}(u|\boldsymbol{y}_i). \end{aligned}$$

Then, at the M-step we maximize the expected value of $\ell^*(\boldsymbol{\theta})$ obtained by substituting each variable a_{juy} with $\hat{a}_{juy}$ and b_u with $\hat{b}_u$. Under the basic LC model, this maximization is simple to perform, as the following explicit solutions are available.

$$\begin{aligned}
\pi_u &= \frac{\hat{b}_u}{n}, \quad u = 1, \ldots, k, \\
\phi_{jy|u} &= \frac{\hat{a}_{juy}}{\hat{b}_u}, \quad j = 1, \ldots, r,\; u = 1, \ldots, k,\; y = 0, \ldots, c-1.
\end{aligned}$$

In the case of individual covariates or constrained LC models, the E-step is performed along the same lines as above. The only difference is that the implementation of the M-step usually requires algorithms for maximizing the weighted log-likelihood of a multinomial logistic or proportional odds model.

Finally, even with reference to the LC model, the general considerations made in Section 2.3 hold about the features of the EM algorithm. In particular, since the likelihood function may be multimodal, different starting values need to be used, taking as the maximum likelihood estimate of the parameters the solution that, at convergence, corresponds to the highest log-likelihood. More details on this point and on different methods to check convergence will be provided in Chapter 3, with reference to the LM model.

2.5.4 Selection of the number of latent classes

A fundamental point in using the LC model, in the basic or advanced versions, is the choice of the number of latent classes, denoted by k. In certain applications, this number is *a priori* defined by the nature of the problem or the particular interest of the user. In most cases, however, it is selected on the basis of the observed data. For this aim, two main approaches are commonly used, which are based on likelihood ratio testing and on information criteria. Other approaches, such as those based on indices that measure the quality of the classification through the posterior probabilities, are more rarely used.

The first approach is based on performing a likelihood ratio test between the model with k classes and that with $k+1$ classes for increasing values of k, until the test is not rejected. The adopted test statistic may be expressed as

$$LR = -2(\hat{\ell}_0 - \hat{\ell}_1), \tag{2.16}$$

where $\hat{\ell}_0$ is the maximum log-likelihood of the smaller model and $\hat{\ell}_1$ is that of the larger model. The problem of this approach is that, in order to obtain a p-value for LR, we need to use a parametric bootstrap procedure (Feng and McCulloch, 1996) based on a suitable number of samples simulated from the estimated model with k classes. This is because the standard regularity conditions, required to validly use the chi-squared distribution to compute these p-values, are not met in this case.

The second approach is based on information criteria. Such criteria are based on indices that are, essentially, penalized versions of the maximum log-likelihood. The two most common criteria of this type are the Akaike information criterion[1] (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978). In the present context, the indices involved in these two criteria may be expressed as follows:

$$\begin{aligned} AIC &= -2\hat{\ell} + 2\#\text{par}, \\ BIC &= -2\hat{\ell} + \log(n)\#\text{par}, \end{aligned}$$

where $\hat{\ell}$ is the maximum log-likelihood of the model and #par denotes the number of parameters of the LC model of interest. The model to be selected is the one with the smallest AIC or BIC. Usually, the second criterion leads to selecting a smaller number of latent classes than the first one, since it is based on a more severe penalization. This difference may be relevant when the sample size is large.

Several studies about the performance of the different approaches for choosing the number of latent classes exist in the literature. In particular, we refer to Dias (2006) for a study referred to the LC model with binary response variables. The problem is strongly related to that of the selection of the number of components of a finite mixture model; then, consider also McLachlan and Peel (2000, Chapter 8). From these studies, it emerges that BIC is usually a trustable criterion to choose the number of latent classes, whereas AIC may lead to a larger number of classes than necessary. We stress that this is not necessarily an undesirable feature of AIC when the latent variable is not of main interest and it is used only to remove the bias due to unobserved heterogeneity in estimating the effect of certain covariates. Finally, the use of the likelihood ratio approach is discouraged by the need of a bootstrap resampling procedure.

Note that, once the number of latent classes has been selected, the same criteria, applied with the same principles as above, may be used to select a reduced model. By reduced model we mean a model with a constrained parametrization, as will be detailed later in the book. However, when applied in this way, the likelihood ratio statistic LR typically has a chi-squared null asymptotic distribution. The number of degrees of freedom of this distribution is equal to the difference between the number of parameters of the larger and the smaller models. With reference to the LM model, a more detailed description of the use of the likelihood ratio testing procedure will be given in Chapter 4.

[1] This criterion was initially named as "An information criterion" by the author who proposed it.

2.6 Applications

In the following we provide some examples based on two datasets among those illustrated in Section 1.4.

2.6.1 Marijuana consumption dataset

We here show the results of the application of the LC model, in the basic version and in a constrained version, to the marijuana consumption dataset illustrated in Section 1.4.1.

In applying the basic LC model to the dataset, we chose $k = 3$ latent classes, which corresponds to the number of categories of the response variables. The model with this number of latent classes will be denoted by M_1; its maximum log-likelihood is equal to $\hat{\ell} = -658.24$ with 32 free parameters. The corresponding values of AIC and BIC are 1380.48 and 1491.45, respectively. The estimates of the conditional response probabilities are reported in Table 2.1, whereas the estimated class weights are given in Table 2.2.

TABLE 2.1
Estimates of the conditional response probabilities $\phi_{jy|u}$ under model M_1

			$\hat{\phi}_{jy\|u}$	
j	u	$y = 0$	$y = 1$	$y = 2$
1	1	0.9732	0.0199	0.0068
	2	0.9401	0.0599	0.0000
	3	0.6959	0.2030	0.1011
2	1	0.9913	0.0087	0.0000
	2	0.6762	0.2523	0.0716
	3	0.3874	0.3256	0.2871
3	1	1.0000	0.0000	0.0000
	2	0.2840	0.6024	0.1137
	3	0.1522	0.2610	0.5868
4	1	0.9414	0.0374	0.0212
	2	0.3548	0.6452	0.0000
	3	0.0000	0.0673	0.9328
5	1	0.8245	0.1251	0.0504
	2	0.3171	0.5866	0.0962
	3	0.0265	0.0960	0.8775

We observe that the latent classes correspond to increasing tendency of marijuana use and have considerably different sizes. In particular, the first class, which is the largest with about 62% of subjects, corresponds to the lowest tendency to this use. Moreover, for the second class, including

TABLE 2.2
Estimates of the initial probabilities π_u under model M_1

u	$\hat{\pi}_u$
1	0.6182
2	0.2149
3	0.1669

about 21% of subjects, we have an intermediate tendency of marijuana consumption, and for the third class, including about 17% of subjects, we have the highest tendency. In fact, when u goes from 1 to 3, the conditional probability of the first category ($y = 0$) tends to decrease for every time occasion, whereas that of the last category ($y = 2$) tends to increase.

In order to better understand the evolution of the phenomena across time, we can use a constrained LC model based on the following assumption concerning the conditional distribution of the response variables given the latent variable

$$\begin{aligned}\log \frac{\phi_{j1|u} + \phi_{j2|u}}{\phi_{j0|u}} &= \alpha_u + \psi_{j1},\\ \log \frac{\phi_{j2|u}}{\phi_{j0|u} + \phi_{j1|u}} &= \alpha_u + \psi_{j2},\end{aligned}$$

for $j = 1, \dots, r$ and $u = 1, \dots, k$, which is based on global logits already defined in Example 3. In particular, the parameters α_u measure the tendency to marijuana consumption for subjects in class u, whereas the parameters ψ_{j1} and ψ_{j2} measure the overall tendency at occasion j. One constraint such as $\alpha_1 = 0$ is necessary to ensure identifiability.

The model based on the above assumption, with the class weights left unrestricted, is denoted by M_2; it has a maximum log-likelihood of -680.38 with 14 parameters and AIC and BIC indices are equal to 1388.8 and 1437.3, respectively. The deviance with respect to model M_1 considered above is

$$LR = -2(-680.38 + 658.24) = 44.28$$

with $32 - 14 = 18$ degrees of freedom and a p-value smaller than 0.001. Then, model M_2 must be rejected in comparison to model M_1; however, it has a smaller BIC and, since it is a useful comparison for the models that we will illustrate in the following, we show in Tables 2.3 and 2.4 its parameter estimates.

We observe that, for each given y, the estimate of ψ_{jy} increases with j, reflecting an increasing tendency of marijuana consumption; for instance,

TABLE 2.3
Estimates of the parameters α_u and ψ_{jy} affecting the conditional response probabilities, given the latent class, under model M_2

u	$\hat{\alpha}_u$	j	y	$\hat{\psi}_{jy}$
1	0.0000	1	1	-6.1779
2	3.2742		2	-7.8474
3	6.0248	2	1	-4.6116
			2	-6.5182
		3	1	-3.2508
			2	-5.4365
		4	1	-2.8277
			2	-4.7676
		5	1	-2.1658
			2	-4.3279

TABLE 2.4
Estimates of the initial probabilities π_u under model M_2

u	$\hat{\pi}_u$
1	0.5678
2	0.2930
3	0.1391

for $y = 1$ the estimate of this parameter is equal to -6.178 for the first time occasion, to -4.612 for the second and so on; these estimates are directly related to the probability that the marijuana consumption is at least once in a month in the past year. Moreover, given the pattern of the estimates of the parameters α_u, the latent classes are still ordered as under model M_1 and the first and third latent classes have a slightly smaller size.

A final comment is on how the above model accounts for the evolution of the phenomena under study. Here, each subject always belongs to the same latent class during all of the period of observation, but the conditional distribution of the response variables is allowed to change as in a latent curve model (see Section 1.3).

On the other hand, LM models have a different perspective since they allow subjects to change latent class but, typically, the interpretation of these classes does not change. In our opinion, this second approach is more interesting for studying how the behavior of a subject evolves during the period of observation.

2.6.2 Criminal conviction history dataset

In order to simplify the illustration, we use only one binary response variable, which is repeatedly observed and indicates convictions for theft and handling stolen goods; see Section 1.4.2. This response variable is denoted by Y_j when referred to age band j, with $j = 1, \dots, r$, where $r = 6$. Moreover, we consider one covariate which is *gender*. This covariate is denoted by X and is equal to 0 for a male and 1 for a female.

For the application to these data, we formulate a model in which the available covariate affects only the latent model. Moreover, we assume that the first latent class corresponds to subjects who have never been convicted for the crime and then we let

$$\phi_{j0|1} = 1,\ \phi_{j1|1} = 0, \quad j = 1, \dots, r.$$

The way in which the covariate affects the probability of the latent classes is based on the following multinomial logit parametrization:

$$\log \frac{\pi_{u|x}}{\pi_{1|x}} = \gamma_{1u} + x\gamma_{2u}, \quad u = 2, \dots, k, \tag{2.17}$$

which is based on comparing the probability of each class (apart from the first) with the first class of never convicted subjects. Note that this assumption is equivalent to the assumption that there are two distinct sets of probabilities for males and females. However, adopting parametrization (2.17) allows us to directly measure the effect of gender through the parameters $\gamma_{22}, \dots, \gamma_{2k}$.

We fitted the model above by the maximum likelihood method for a different number of latent classes. Note that, since we express the model log-likelihood as in (2.9), we can include the number of males or females who have never committed a crime in the period of observation. The results of this preliminary fitting, in terms of maximum log-likelihood, number of parameters, and corresponding values of AIC and BIC, are reported in Table 2.5. We fitted the model increasing the number of classes until $k = 9$, since this is the first case in which the value of AIC is greater than that for the previous number of classes.

The sequence of AIC values in Table 2.5 leads to selecting $k = 8$ classes, whereas the sequence of BIC values leads to selecting $k = 5$ classes. However, we prefer to rely on this second choice, as BIC is considered a more reliable criterion. The corresponding model is denoted by M_1.

With $k = 5$, we obtain the parameter estimates displayed in Tables 2.6 and 2.7. In reporting these estimates, the latent classes are ordered according to the probability of being convicted for the crime at the first occasion. Note that, since the response variables are binary, in the first of

TABLE 2.5
Maximum log-likelihood, number of parameters, and AIC and BIC indices for a number of latent classes between 1 and 9

k	$\hat{\ell}$	#par	AIC	BIC
1	-46715.15	6	93442.30	93495.53
2	-42176.38	8	84368.76	84439.74
3	-41282.03	16	82596.05	82738.00
4	-41068.45	24	82184.89	82397.81
5	-40982.21	32	82028.43	82312.32
6	-40962.01	40	82004.02	82358.88
7	-40944.39	48	81984.77	82410.60
8	-40934.15	56	81980.30	82477.11
9	-40926.66	64	81981.31	82549.09

these tables we report only the estimates of the probabilities of "success," that is, the estimates of the parameters $\phi_{j1|u}$.

On the basis of results in Table 2.6 we can characterize each latent class in terms of propensity to commit a crime and the evolution of this

TABLE 2.6
Estimates of the conditional response probabilities $\phi_{j1|u}$ under model M_1

	$\hat{\phi}_{j1\|u}$				
j	$u=1$	$u=2$	$u=3$	$u=4$	$u=5$
1	0.0000	0.0198	0.0703	0.4033	0.4394
2	0.0000	0.0298	0.1099	0.6278	0.7653
3	0.0000	0.0821	0.0463	0.3738	0.8068
4	0.0000	0.1362	0.0000	0.2001	0.8110
5	0.0000	0.0894	0.0006	0.0832	0.6368
6	0.0000	0.0451	0.0001	0.0330	0.3593

TABLE 2.7
Estimates of the parameters γ_{1u} and γ_{2u} under model M_1

u	$\hat{\gamma}_{1u}$	$\hat{\gamma}_{2u}$
2	-0.7421	-1.3549
3	0.4273	-1.9990
4	-1.4313	-6.5513
5	-3.1114	-3.3702

propensity. In particular, we easily observe that subjects in the last class have the highest propensity to commit a crime, contrary to subjects in the first class who have no propensity at all to commit a crime (never offenders). For instance, subjects in the last class have a probability equal to 0.4394 to commit a crime in the first period, equal to 0.7653 to commit a crime in the second period, and so on. The other classes are not strictly ordered in terms of conditional probability of being convicted at each time occasion, since these probabilities do not follow a unique trend in time. Then these classes cannot be so easily characterized.

The results in Table 2.7 lead us to conclude that the covariate gender has a negative effect on the logit of the probability of being in any latent class, apart from the first. In other words, females have a higher probability, with respect to males, of being never offenders. This is confirmed by the two vectors of probabilities of U which are reported, for males and females, in Table 2.8. In particular, we have that around 30% of males are never offenders as opposed to around 75% of females.

TABLE 2.8
Estimated distribution of U for males and females under model M_1

u	$\hat{\pi}_{u\|0}$	$\hat{\pi}_{u\|1}$
1	0.3037	0.7505
2	0.1446	0.0922
3	0.4656	0.1559
4	0.0726	0.0003
5	0.0135	0.0011

A final point is whether it is really important to include the covariate gender in the model. For this aim, we also fitted the model without this covariate and with the same number of latent classes; this model is denoted by M_2. It has a maximum log-likelihood of -42628.00 with 28 parameters, and AIC and BIC indices are equal to 85312.00 and 85560.40, respectively. Moreover, the deviance between models M_1 and M_2 is $LR = 3291.57$, which must be compared with a chi-squared distribution with 4 degrees of freedom. These results undoubtedly lead us to conclude that gender is a significant covariate.

2.7 Markov chain model for longitudinal data

The MC model represents a fundamental model for stochastic processes. In the following, we briefly review this model in the context of longitudinal data, where it is also referred to as *transition model*. We introduce only the concepts which are necessary for the main topic of the book, whereas for an overall treatment on Markov chains, we suggest standard books on stochastic processes, such as Grimmett and Stirzaker (2001) and Taylor and Karlin (1998). For a review in the context of longitudinal data, see also Frees (2004).

As for the LC model, we first review the basic version of the model, then more advanced versions, and finally maximum likelihood estimation.

2.7.1 Basic version

Since we are explicitly referring to the context of longitudinal data, we use a slightly different notation for the response variables, which makes explicit their dependence on time. In particular, suppose that each subject is observed at T consecutive time occasions. We let $Y^{(1)},\dots,Y^{(T)}$ denote the corresponding response variables, which are categorical with c categories indexed from 0 to $c-1$. As before, these variables are collected in the vector $\boldsymbol{Y}$. The same notation will be used in the following chapters.

The main assumption of the MC model of *order* o is that $Y^{(t)}$ is conditionally independent of $Y^{(1)},\dots,Y^{(t-o-1)}$ given $Y^{(t-o)},\dots,Y^{(t-1)}$, that is,

$$\begin{aligned}&f_{Y^{(t)}|Y^{(1)},\dots,Y^{(t-1)}}(y^{(t)}|y^{(1)},\dots,y^{(t-1)})\\&=f_{Y^{(t)}|Y^{(t-o)},\dots,Y^{(t-1)}}(y^{(t)}|y^{(t-o)},\dots,y^{(t-1)})\end{aligned} \tag{2.18}$$

for $t=o+1,\dots,T$ and any configuration of $Y^{(1)},\dots,Y^{(t)}$. When $o=0$, an independence model results, whereas when $o=1$, we have a *first-order MC model*, which leads to a strong simplification of the dependence structure between the response variables. In fact, as shown by the path diagram in Figure 2.2, each response variable $Y^{(t)}$ is directly affected only by the lagged response variable. This is the main case in most applications.

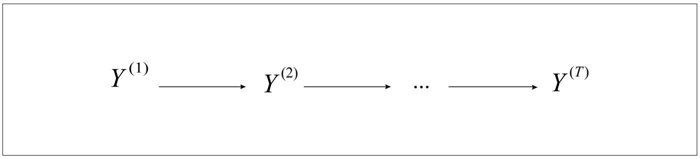

FIGURE 2.2
Path diagram for the first-order MC model

Under assumption (2.18), the distribution of $\boldsymbol{Y}$ has probability mass function

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \prod_{t=1}^{o} f_{Y^{(t)}|Y^{(1)},\dots,Y^{(t-1)}}(y^{(t)}|y^{(1)},\dots,y^{(t-1)}) \\ &\times \prod_{t=o+1}^{T} f_{Y^{(t)}|Y^{(t-o)},\dots,Y^{(t-1)}}(y^{(t)}|y^{(t-o)},\dots,y^{(t-1)}), \end{aligned}$$

for any possible realization $\boldsymbol{y}$ of $\boldsymbol{Y}$. In particular, for the first-order MC model we have

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \pi_{y^{(1)}} \prod_{t=2}^{T} \pi^{(t)}_{y^{(t)}|y^{(t-1)}}, \tag{2.19}$$

where we use the notation

$$\pi_y = f_{Y^{(1)}}(y), \quad y = 0,\dots,c-1,$$

for the *initial probabilities* and

$$\pi^{(t)}_{y|\bar{y}} = f_{Y^{(t)}|Y^{(t-1)}}(y|\bar{y}), \quad t = 2,\dots,T, \quad \bar{y}, y = 0,\dots,c-1,$$

for the *transition probabilities*. In the above expressions we use y to denote a realization of $Y^{(t)}$ and $\bar{y}$ for a realization of $Y^{(t-1)}$.

An expression similar to (2.19) may be used to obtain the probability mass function of $Y^{(1)},\dots,Y^{(t)}$ for $t = 2,\dots,T$. For a first-order model, in particular, we have

$$f_{Y^{(1)},\dots,Y^{(t)}}(y^{(1)},\dots,y^{(t)}) = \pi_{y^{(1)}} \prod_{s=2}^{t} \pi^{(s)}_{y^{(s)}|y^{(s-1)}}. \tag{2.20}$$

Then, by marginalization, we obtain the distribution of a single response variable, that is,

$$f_{Y^{(t)}}(y^{(t)}) = \sum_{y^{(1)}=0}^{c-1} \cdots \sum_{y^{(t-1)}=0}^{c-1} f_{Y^{(1)},\dots,Y^{(t)}}(y^{(1)},\dots,y^{(t)}).$$

Under the first-order *basic MC model*, the initial and transition probabilities are the parameters to estimate. These parameters must satisfy the usual constraint of nonnegativity, further to the constraints $\sum_{y=0}^{c-1} \pi_y = 1$ and $\sum_{y=0}^{c-1} \pi_{y|\bar{y}}^{(t)} = 1$, $\bar{y} = 0, \ldots, c-1$, $t = 2, \ldots, T$. Then, the number of free parameters is

$$\#\text{par} = \underbrace{c-1}_{\pi_y} + \underbrace{(T-1)c(c-1)}_{\pi_{y|\bar{y}}^{(t)}}.$$

Similar constraints are necessary in the case of higher-order Markov models, and the number of parameters are easily computable even in these cases.

2.7.2 Advanced versions

The parameters of the basic MC model may be constrained in order to make it more parsimonious and easily interpretable. The most natural constraint is that of *time homogeneity*, according to which the transition probabilities do not depend on t. In the relevant case of first-order models, we have

$$\pi_{y|\bar{y}}^{(t)} = \pi_{y|\bar{y}}, \quad t = 2, \ldots, T, \ \bar{y}, y = 0, \ldots, c-1, \tag{2.21}$$

where $\pi_{y|\bar{y}}$ are transition probabilities common to all time occasions.

A greater variety of constraints may be formulated by suitable parametrizations based on logits or similar effects, as we show in the following example.

Example 5 — Proportional odds model on the transition probabilities. *In the case of ordinal response variables and first-order dependence structure, consider the following parametrization:*

$$\log \frac{\pi_{y|\bar{y}}^{(t)} + \cdots + \pi_{c-1|\bar{y}}^{(t)}}{\pi_{0|\bar{y}}^{(t)} + \cdots + \pi_{y-1|\bar{y}}^{(t)}} = \gamma_{1\bar{y}} + \gamma_{2y}, \quad y = 1, \ldots, c-1,$$

for $t = 2, \ldots, T$ and $\bar{y} = 0, \ldots, c-1$, which is based on global logits on the transition probabilities. One identifiability constraint, such as $\gamma_{10} = 0$, is required.

Note that the above parametrization incorporates the constraint of time homogeneity in (2.21) and strongly reduces the number of parameters that becomes equal to $2(c-1)$, with the parameters $\gamma_{1\bar{y}}$ measuring the effect of the previous state on the conditional distribution of the new state.

In the presence of individual covariates, which are collected in the column vector $\boldsymbol{X}$, these covariates may be used to suitably parameterize the initial and transition probabilities. Moreover, we may formulate the MC model by combining parametrizations depending on individual covariates with constrained parametrizations, as in the following example. In this case, we adopt the following notation for the initial probabilities:

$$\pi_{y|\boldsymbol{x}} = f_{Y^{(1)}|\boldsymbol{X}}(y|\boldsymbol{x}), \quad y = 0, \dots, c-1,$$

and the following notation for the transition probabilities:

$$\pi^{(t)}_{y|\boldsymbol{x}\bar{y}} = f_{Y^{(t)}|\boldsymbol{X},Y^{(t-1)}}(y|\boldsymbol{x}, \bar{y}), \quad t = 2, \dots, T,\ \bar{y}, y = 0, \dots, c-1,$$

where $\boldsymbol{x}$ is a configuration of $\boldsymbol{X}$.

Example 6 — Proportional odds model on the initial probabilities. *Again in the case of ordinal response variables, and under a first-order dependence structure, assume that*

$$\log \frac{\pi_{y|\boldsymbol{x}} + \cdots + \pi_{c-1|\boldsymbol{x}}}{\pi_{0|\boldsymbol{x}} + \cdots + \pi_{y-1|\boldsymbol{x}}} = \gamma_{1y} + \boldsymbol{x}'\boldsymbol{\gamma}_2, \quad y = 1, \dots, c-1.$$

This is a proportional-odds model on the initial probabilities that may be assumed jointly with the constraint

$$\pi^{(t)}_{y|\boldsymbol{x}\bar{y}} = \pi_{y|\bar{y}}, \quad t = 2, \dots, T,\ \bar{y}, y = 0, \dots, c-1,$$

for all possible covariate configurations $\boldsymbol{x}$*; this incorporates the constraint of time homogeneity. The resulting model then assumes the same evolution for all subjects and time occasions.*

Above we presented some examples of models that may be formulated on the basis of a Markov chain approach. A systematic review of constraints on initial and transition probabilities and related methods to include individual covariates will be given in Chapters 4 and 5 with reference to LM models.

2.7.3 Likelihood inference

In the absence of individual covariates, let $\boldsymbol{y}_i$, $i = 1, \dots, n$, denote the observed response configurations for the sample units and let $n_{\boldsymbol{y}}$ denote the frequency of the configuration $\boldsymbol{y}$. Then, the likelihood has the usual expression

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{Y}}(\boldsymbol{y}_i) = \sum_{\boldsymbol{y}} n_{\boldsymbol{y}} \log f_{\boldsymbol{Y}}(\boldsymbol{y}),$$

where $\boldsymbol{\theta}$ is the vector of all model parameters. With individual covariates, collected in the vectors $\boldsymbol{x}_i$, $i = 1, \ldots, n$, let $n_{\boldsymbol{xy}}$ be the frequency of the covariate configuration $\boldsymbol{x}$ and the response configuration $\boldsymbol{y}$. Then, the log-likelihood becomes

$$\ell(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}} n_{\boldsymbol{xy}} \log f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}),$$

where the sum is over all covariate and response configurations observed at least once. In any case we prefer to use the previous definition.

In the following, we discuss the maximization of this log-likelihood, so as to estimate $\boldsymbol{\theta}$, and briefly review model selection.

2.7.4 Maximum likelihood estimation

For the basic MC model, explicit formulae are available for the maximum likelihood estimation of the parameters. In order to show this, we consider a first-order model. In this case, the log-likelihood may be simplified as

$$\ell(\boldsymbol{\theta}) = \sum_{y=0}^{c-1} a_y^{(1)} \log \pi_y + \sum_{t=2}^{T} \sum_{\bar{y}=0}^{c-1} \sum_{y=0}^{c-1} a_{\bar{y}y}^{(t)} \log \pi_{y|\bar{y}}^{(t)},$$

where $a_y^{(t)}$ is the number of units having response equal to y at occasion t and $a_{\bar{y}y}^{(t)}$ is the number of subjects responding by $\bar{y}$ at occasion $t-1$ and by y at occasion t.

After some simple algebra, we can easily realize that $\ell(\boldsymbol{\theta})$ is maximized by the following parameter values

$$\begin{aligned}
\hat{\pi}_y &= \frac{a_y^{(1)}}{n}, \quad y = 0, \ldots, c-1, \\
\hat{\pi}_{y|\bar{y}}^{(t)} &= \frac{a_{\bar{y}y}^{(t)}}{a_{\bar{y}}^{(t-1)}}, \quad t = 2, \ldots, T,\ \bar{y}, y = 0, \ldots, c-1.
\end{aligned}$$

Then, these are the maximum likelihood estimates of the parameters. Simple formulae may be used for higher-order MC models.

More complex models typically require an iterative algorithm, such as the NR or the Fisher-scoring, to maximize this log-likelihood. These algorithms are based on the first derivative vector of $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and on its second derivative or the expected information matrix. Since the same algorithms will be used for the estimation of certain LM models, the reader is referred to the remaining chapters, in particular Chapters 4 and 5, for a detailed description.

Concerning standard errors to be associated to the parameter estimates, the same considerations as in Section 2.4 hold for the MC model. Given the introductory level of this chapter, even for this model we do not enter into details about standard errors and we skip their computation in the following applications.

2.7.5 Model selection

The first characteristic that needs to be chosen in applying an MC model is the order o. In this regard, we can use the same criteria illustrated in Section 2.5.4 to select the number of latent classes of the LC model. In particular, the information criteria AIC and BIC are typically used.

We have to clarify that the same criteria mentioned above are typically used to select a model based on constraints on the initial and transition probabilities or based on suitable parametrizations of these probabilities which depend on individual covariates. In this regard, we can also use the likelihood ratio statistic LR, defined in (2.16), which has a null asymptotic distribution of chi-squared type in regular cases.

2.8 Applications

In this section we illustrate the MC model by reanalyzing the two datasets about marijuana consumption and criminal conviction histories.

2.8.1 Marijuana consumption dataset

For the marijuana consumption dataset, the basic first-order MC model, denoted by M_3, has maximum log-likelihood equal to -655.16 with 26 parameters, so that AIC and BIC indices are equal to 1362.32 and 1452.49, respectively. The parameter estimates under this model are reported in Tables 2.9 and 2.10.

These estimates indicate that most subjects are nonconsumers at the beginning of the period of observation. Nonconsumers always have a probability close to 0.8 of remaining nonconsumers from wave to wave. The other categories have a lower persistence, in particular the second. In order to make clearer the interpretations of these results, we can also consider the estimated marginal distribution computed on the basis of (2.20). The results are reported in Table 2.11.

TABLE 2.9
Estimates of the initial probabilities π_y under model M_3

y	$\hat{\pi}_y$
0	0.9198
1	0.0591
2	0.0211

TABLE 2.10
Estimates of the transition probabilities $\pi^{(t)}_{y|\bar{y}}$ under model M_3

		$\hat{\pi}^{(t)}_{y\|\bar{y}}$		
t	$\bar{y}$	$y = 0$	$y = 1$	$y = 2$
2	0	0.8624	0.1101	0.0275
	1	0.4286	0.1429	0.4286
	2	0.2000	0.2000	0.6000
3	0	0.8256	0.1436	0.0308
	1	0.1852	0.3704	0.4444
	2	0.0667	0.2000	0.7333
4	0	0.8623	0.0838	0.0539
	1	0.2439	0.5122	0.2439
	2	0.0690	0.2069	0.7241
5	0	0.7756	0.1667	0.0577
	1	0.3415	0.5366	0.1220
	2	0.0750	0.1000	0.8250

TABLE 2.11
Estimated marginal distribution of the response variable for each time occasion under model M_3

	$\hat{f}_{Y^{(t)}}(y)$		
t	$y = 0$	$y = 1$	$y = 2$
1	0.9198	0.0591	0.0211
2	0.8228	0.1139	0.0633
3	0.7046	0.1730	0.1224
4	0.6582	0.1730	0.1688
5	0.5823	0.2194	0.1983

These results show that the tendency to use marijuana increases during the period of observation, as the probability in the first category constantly decreases in favor of that in the second and third categories.

We also consider the time-homogeneous model based on constraint

(2.21). This new model, denoted by M_4, has a maximum log-likelihood equal to -669.37 with 8 parameters, and AIC and BIC indices are equal to 1354.74 and 1382.49, respectively. The likelihood ratio test statistic between this model and M_3 is then 28.42. The asymptotic p-value for this statistic, based on a chi-squared distribution with 18 degrees of freedom, is 0.056. Overall, model M_4 is preferable to model M_3. The estimates of its parameters are reported in Table 2.12 about the transition probabilities, whereas we have the same estimated initial probabilities as in Table 2.9.

TABLE 2.12
Estimates of the transition probabilities $\pi_{y|\bar{y}}$ under model M_4

	$\hat{\pi}_{y\|\bar{y}}$		
$\bar{y}$	$y = 0$	$y = 1$	$y = 2$
1	0.8342	0.1250	0.0408
2	0.2846	0.4472	0.2683
3	0.0787	0.1573	0.7640

Note that the matrix is rather close to symmetry, implying that the probability of increasing marijuana consumption is close to that of decreasing. However, considering that most subjects do not use marijuana at the beginning of the study, the overall tendency to consumption increases during the period of observation as a larger number of subjects moves from category 0 to categories 1 and 2 than vice versa. This conclusion is an agreement with the structure of the marginal distributions in Table 2.11.

2.8.2 Criminal conviction history dataset

We now consider the analysis, by the MC model, of the dataset about conviction histories. Along the same lines as in Section 2.6.2, we consider the response variable referring to conviction for theft and handling stolen goods. We also consider the covariate *gender*, denoted by X, which is equal to 0 for a male and 1 for a female.

The model here used for the analysis of the data, denoted by M_3, is a first-order MC model based on the following parametrization of the initial probabilities

$$\log \frac{\pi_{1|x}}{\pi_{0|x}} = \gamma_1 + x\gamma_2, \quad x = 0, 1, \tag{2.22}$$

where the parameter γ_2 measures the effect of the gender on the initial probabilities. The following parametrization is assumed on the transition probabilities

$$\log \frac{\pi_{1|x\bar{y}}^{(t)}}{\pi_{0|x\bar{y}}^{(t)}} = \delta_{1\bar{y}} + \delta_{2t} + x\delta_3, \quad t = 2, \ldots, T, \ x, \bar{y} = 0, 1,$$

so that $\delta_{1\bar{y}}$ is an intercept specific for the previous value of the response variable, δ_{2t} are parameters taking the time effect into account (in order to ensure identifiability we let $\delta_{22} = 0$), and δ_3 measures the effect of gender on the transition probabilities.

The MC model based on the above assumptions has maximum log-likelihood equal to -41756.29 with 9 parameters, so that AIC and BIC indices are equal to 83530.58 and 83610.42, respectively. We also have the parameter estimates in Table 2.13.

TABLE 2.13
Estimates of the parameters under model M_3

$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\bar{y}$	$\hat{\delta}_{1\bar{y}}$	t	$\hat{\delta}_{2t}$	$\hat{\delta}_3$
-2.5641	-1.7890	0	-2.4256	3	-0.6265	-1.2538
		1	-0.2856	4	-0.9297	
				5	-1.3209	
				6	-1.9630	

According to these estimates, females have a smaller tendency than males to be offenders at the beginning of the period of observation (because $\hat{\gamma}_2 = -1.7890$) and to move from state 0 to state 1 (because $\hat{\delta}_3 = -1.2538$). Note that assumption (2.22) is equivalent to the assumption that males and females have different vectors of initial probabilities, the estimates of which are reported in Table 2.14.

TABLE 2.14
Estimated distribution of Y_1 for males and females under model M_3

y	$\hat{\pi}_{y\|0}$	$\hat{\pi}_{y\|1}$
0	0.9285	0.9873
1	0.0715	0.0127

Finally, in order to understand if the covariate is significant in predicting the response variable, we estimated model M_4 in which this covariate does not affect the initial probabilities, and then $\gamma_2 = 0$. We also estimated model M_5 in which the covariate does not affect the transition probabilities, and then $\delta_3 = 0$. The maximum log-likelihood is equal to -42370.83 for model M_4 and -42917.74 for model M_5. Both values are much lower than the value of the maximum log-likelihood of model M_3 with only one less parameter, and then the hypotheses behind these models must be rejected.

3

Basic latent Markov model

3.1 Introduction

In this chapter, we illustrate the *basic formulation* of the latent Markov (LM) model for categorical response variables; this formulation assumes the same measurement model, which is also time homogenous, and the same latent model for all observational units. Therefore, individual covariates are ruled out. At the same time, no constraints are posed either on the conditional distribution of each occasion-specific response variable, given the corresponding latent state, or on the initial and transition probabilities. As we will show, the resulting model is rather easy to estimate by the Expectation–Maximization (EM) algorithm.

In presenting the model, we disentangle the case of univariate responses from that of multivariate responses. In the first case, we observe only one response variable at each occasion, whereas, in the second case, we observe more response variables at each occasion. We also explicitly consider *balanced panel* data, where all units are observed at the same number of time occasions. However, the proposed methodology may be easily extended to the case of unbalanced panel data, where not always the same number of observations is available for each unit and there are missing responses. This extension will be dealt with in Chapter 7.

3.2 Univariate formulation

In the case of univariate data, for every sample unit we observe one response variable, denoted by $Y^{(t)}$, for each time occasion t, with $t = 1, \ldots, T$. The vector with elements $Y^{(1)}, \ldots, Y^{(T)}$ is denoted by $\tilde{\boldsymbol{Y}}$; as will be clear in the following, the need to introduce the tilde over $\boldsymbol{Y}$ is to avoid confusion with the multivariate case. Typically, the response variables have the same nature, as they correspond to repeated measurements on the same subjects at different occasions, and then we denote by c the number of their categories, coded from 0 to $c-1$. However, the approach may also be applied to the case of response variables having a different nature and, possibly, a different number of categories.

In the above framework, the main assumption of the basic LM model is that the response variables in $\tilde{\boldsymbol{Y}}$ are conditionally independent given the latent process $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(T)})$. This assumption is a form of *local independence*, already discussed in Section 2.2. The latent process is assumed to follow a first-order Markov chain with state space $\{1, \ldots, k\}$, where k is the number of latent states. Then, for $t = 2, \ldots, T$, the latent variable $U^{(t)}$ is conditionally independent of $U^{(1)}, \ldots, U^{(t-2)}$ given $U^{(t-1)}$. See Figure 3.1 for an illustration via path diagram.

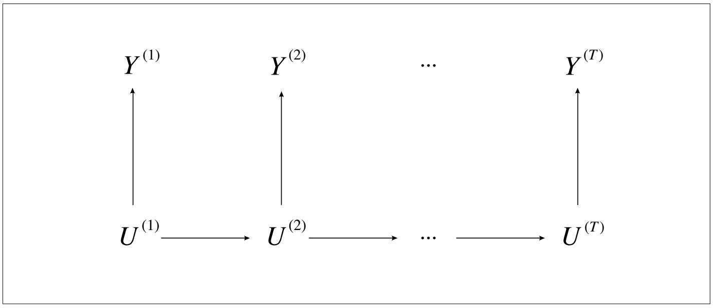

FIGURE 3.1
Path diagram for the basic LM model for univariate data

In order to interpret the model based on the above assumptions, it is useful to consider that it may be seen as a generalization of a standard Markov chain (MC) model (see Section 2.7) to account for *measurement errors*. In particular, the outcome $U^{(t)}$ is observed with measurement error as $Y^{(t)}$, and then it is reasonable to assume that the number of observable categories (c) is equal to that of latent states (k). For a discussion on this point see Wiggins (1973, Chapter 4). Furthermore, as already mentioned, the LM model may be seen as a generalization of a latent class (LC) model in which each subject may move between latent classes. This interpretation is clarified by comparing Figure 3.1 with Figure 2.1. In this case, k represents the number of subpopulations (or latent classes) of interest and is not strictly related to c. Overall, the first-order Markov assumption for the latent process is easily interpretable, and it is seldom found to be restrictive.

In order to better understand the model structure, it is worth summarizing its parameters. These parameters are the conditional response probabilities

$$\phi_{y|u} = f_{Y^{(t)}|U^{(t)}}(y|u), \quad t = 1,\dots,T,\ u = 1,\dots,k,\ y = 0,\dots,c-1,$$

the initial probabilities

$$\pi_u = f_{U^{(1)}}(u), \quad u = 1,\dots,k,$$

and the transition probabilities

$$\pi_{u|\bar{u}}^{(t)} = f_{U^{(t)}|U^{(t-1)}}(u|\bar{u}), \quad t = 2,\dots,T,\ \bar{u}, u = 1,\dots,k;$$

in the last expression, u is a realization of $U^{(t)}$, whereas $\bar{u}$ is a realization of $U^{(t-1)}$. Note that all these probabilities do not depend on the specific sample unit since, in its basic version, the LM model does not account for individual covariates. Moreover, due to the usual constraints on the above probabilities, the number of free parameters is

$$\#\text{par} = \underbrace{k(c-1)}_{\phi_{y|u}} + \underbrace{k-1}_{\pi_u} + \underbrace{(T-1)k(k-1)}_{\pi_{u|\bar{u}}^{(t)}}, \tag{3.1}$$

which, in the case of binary response variables, becomes

$$\#\text{par} = 2k + (T-1)k(k-1) - 1.$$

A summary of the model parameters is given in Table 3.1.

On the basis of the above parameters, the probability mass function of the distribution of $\boldsymbol{U}$ may be expressed as

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = \pi_{u^{(1)}} \prod_{t=2}^{T} \pi_{u^{(t)}|u^{(t-1)}}^{(t)}, \tag{3.2}$$

TABLE 3.1
Summary of the parameters of the basic LM model

Parameter	Description	Range
$\phi_{y\|u}$	conditional response probabilities	$u = 1\dots,k$ $y = 0,\dots,c-1$
π_u	initial probabilities of the latent process	$u = 1,\dots,k$
$\pi_{u\|\bar{u}}^{(t)}$	transition probabilities of the latent process	$t = 2,\dots,T$ $\bar{u} = 1,\dots,k$ $u = 1,\dots,k$

where $\boldsymbol{u}$ denotes a realization of $\boldsymbol{U}$, with elements $u^{(1)},\dots,u^{(T)}$. Moreover, about the conditional distribution of $\tilde{\boldsymbol{Y}}$ given $\boldsymbol{U}$, we have

$$f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U}}(\tilde{\boldsymbol{y}}|\boldsymbol{u}) = \prod_{t=1}^{T} \phi_{y^{(t)}|u^{(t)}}, \tag{3.3}$$

for any realization $\tilde{\boldsymbol{y}}$ of $\tilde{\boldsymbol{Y}}$. Finally, equation (2.1) implies that, for the *manifest distribution* of $\tilde{\boldsymbol{Y}}$, we have

$$f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}}) = \sum_{\boldsymbol{u}} \pi_{u^{(1)}} \pi^{(2)}_{u^{(2)}|u^{(1)}} \cdots \pi^{(T)}_{u^{(T)}|u^{(T-1)}} \phi_{y^{(1)}|u^{(1)}} \cdots \phi_{y^{(T)}|u^{(T)}}. \tag{3.4}$$

It is important to note that computing $f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$ as expressed above involves a sum over all the possible k^T configurations of the vector $\boldsymbol{u}$. This typically requires a considerable computational effort; the solution of this problem is discussed below. The following example may be useful to clarify these aspects.

Example 7 — Computation of the manifest distribution. *Suppose that we observe subjects at three occasions ($T = 3$) and that the latent Markov chain is based on three states ($k = 3$). Regardless of the number of categories of the response variables (c), the sum in (3.4) is over all the possible $3^3 = 27$ configurations $\boldsymbol{u} = (u^{(1)}, u^{(2)}, u^{(3)})$. Arranging these configurations in lexicographical order, this sum for $f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$ may be explicitly written as follows for any response configuration $\tilde{\boldsymbol{y}} = (y^{(1)}, y^{(2)}, y^{(3)})$:*

$$\begin{aligned}&\pi_1\pi^{(2)}_{1|1}\pi^{(3)}_{1|1}\phi_{y^{(1)}|1}\phi_{y^{(2)}|1}\phi_{y^{(3)}|1} + \pi_1\pi^{(2)}_{1|1}\pi^{(3)}_{2|1}\phi_{y^{(1)}|1}\phi_{y^{(2)}|1}\phi_{y^{(3)}|2} + \\ &+ \cdots + \pi_3\pi^{(2)}_{3|3}\pi^{(3)}_{3|3}\phi_{y^{(1)}|3}\phi_{y^{(2)}|3}\phi_{y^{(3)}|3}.\end{aligned}$$

Obviously, this sum would involve a much larger number of configurations with a greater T and/or k.

In order to efficiently compute the probability $f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$, we can use a forward recursion (Baum et al., 1970; Welch, 2003) for obtaining

$$q^{(t)}(u, \tilde{\boldsymbol{y}}) = f_{U^{(t)},Y^{(1)},\dots,Y^{(t)}}(u, y^{(1)}, \dots, y^{(t)}), \quad t = 1, \dots, T.$$

Then, we have

$$f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}}) = \sum_{u=1}^{k} q^{(T)}(u, \tilde{\boldsymbol{y}}). \tag{3.5}$$

In particular, given $q^{(t-1)}(u,\tilde{\boldsymbol{y}})$ for $u=1,\dots,k$, the t-th iteration of the recursion, $t=2,\dots,T$, consists of computing

$$q^{(t)}(u,\tilde{\boldsymbol{y}})=\sum_{\bar{u}=1}^{k} q^{(t-1)}(\bar{u},\tilde{\boldsymbol{y}})\pi^{(t)}_{u|\bar{u}}\phi_{y^{(t)}|u},\quad u=1,\dots,k, \tag{3.6}$$

starting with

$$q^{(1)}(u,\tilde{\boldsymbol{y}})=\pi_u\phi_{y^{(1)}|u},\quad u=1,\dots,k. \tag{3.7}$$

Example 8 — Implementation of the forward recursion. *With reference to the same case as in Example 7, the forward recursion is based on sequentially computing*

$$\begin{aligned}
q^{(1)}(u,\tilde{\boldsymbol{y}}) &= \pi_u\phi_{y^{(1)}|u},\quad u=1,\dots,k,\\
q^{(2)}(u,\tilde{\boldsymbol{y}}) &= \sum_{\bar{u}=1}^{k} q^{(1)}(\bar{u},\tilde{\boldsymbol{y}})\pi^{(2)}_{u|\bar{u}}\phi_{y^{(2)}|u},\quad u=1,\dots,k,\\
q^{(3)}(u,\tilde{\boldsymbol{y}}) &= \sum_{\bar{u}=1}^{k} q^{(2)}(\bar{u},\tilde{\boldsymbol{y}})\pi^{(3)}_{u|\bar{u}}\phi_{y^{(3)}|u},\quad u=1,\dots,k.
\end{aligned}$$

Finally, we have the manifest probability

$$f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})=q^{(3)}(1,\tilde{\boldsymbol{y}})+q^{(3)}(2,\tilde{\boldsymbol{y}})+q^{(3)}(3,\tilde{\boldsymbol{y}}).$$

The above recursion may be simply implemented by using the matrix notation (Bartolucci, 2006; Bartolucci et al., 2007). This is shown in detail in Appendix 1.

In certain contexts, the marginal distribution of $U^{(t)}$, $t=1,\dots,T$, is of interest. A similar problem was dealt with in Section 2.7.1. Obviously, the distribution of $U^{(1)}$ has probability mass function with values π_u, $u=1,\dots,k$. On the other hand, in order to obtain the distribution of $U^{(t)}$, for $t=2,\dots,T$, we can first compute $f_{U^{(1)},\dots,U^{(t)}}(u^{(1)},\dots,u^{(t)})$ by an expression that closely recalls expression (3.2), that is,

$$f_{U^{(1)},\dots,U^{(t)}}(u^{(1)},\dots,u^{(t)})=\pi_{u^{(1)}}\prod_{s=2}^{t}\pi^{(s)}_{u^{(s)}|u^{(s-1)}}.$$

Then, $f_{U^{(t)}}$ may be obtained by marginalization as

$$f_{U^{(t)}}(u^{(t)})=\sum_{u^{(1)}=1}^{k}\cdots\sum_{u^{(t-1)}=1}^{k} f_{U^{(1)},\dots,U^{(t)}}(u^{(1)},\dots,u^{(t)}).$$

It is worth noting that the probabilities may be computed by a simplified version of the above recursion. Let $q^{(t)}(u) = f_{U^{(t)}}(u)$; then, we have

$$q^{(t)}(u) = \sum_{\bar{u}=1}^{k} q^{(t-1)}(\bar{u})\pi_{u|\bar{u}}^{(t)}, \quad t = 2, \dots, T, \ u = 1, \dots, k, \tag{3.8}$$

initialized with $q^{(1)}(u) = \pi_u$, $u = 1, \dots, k$.

As will be illustrated in Section 3.7.1, the marginal distribution of each variable $U^{(t)}$ may be suitably plotted in order to give a representation of the evolution of the phenomenon of interest. Moreover, on the basis of the above probabilities we can obtain the marginal distribution of $Y^{(t)}$ as

$$f_{Y^{(t)}}(y) = \sum_{u=1}^{k} f_{U^{(t)}}(u)\phi_{y|u}, \quad y = 0, \dots, c-1.$$

3.3 Multivariate formulation

In the multivariate case, we observe a vector of response variables $\boldsymbol{Y}^{(t)} = (Y_1^{(t)}, \dots, Y_r^{(t)})$ for $t = 1, \dots, T$. Each variable $Y_j^{(t)}$ has c_j levels coded from 0 to $c_j - 1$. However, we still denote by $\tilde{\boldsymbol{Y}}$ the response vector which in this case is made up of the union of the vectors $\boldsymbol{Y}^{(t)}$, $t = 1, \dots, T$. The LM model presented in the previous section may be generalized to this case by considering an extended version of the assumption of local independence. Beyond assuming that the vectors $\boldsymbol{Y}^{(t)}$, $t = 1, \dots, T$, are conditionally independent given $\boldsymbol{U}$, we assume that the response variables in each of these vectors are conditionally independent given $U^{(t)}$. The resulting model is represented by the path diagram shown in Figure 3.2.

Under this formulation, the new parameters are the probabilities $\phi_{jy|u} = f_{Y_j^{(t)}|U^{(t)}}(y|u)$, with $j = 1, \dots, r$, $t = 1, \dots, T$, $u = 1, \dots, k$, and $y = 0, \dots, c_j - 1$. Then, the number of free parameters becomes

$$\#\text{par} = \underbrace{k\sum_{j=1}^{r}(c_j - 1)}_{\phi_{jy|u}} + \underbrace{k-1}_{\pi_u} + \underbrace{(T-1)k(k-1)}_{\pi_{u|\bar{u}}^{(t)}},$$

since we have a larger number of parameters involved in the measurement model. With binary response variables, we have

$$\#\text{par} = k(r+1) + (T-1)k(k-1) - 1.$$

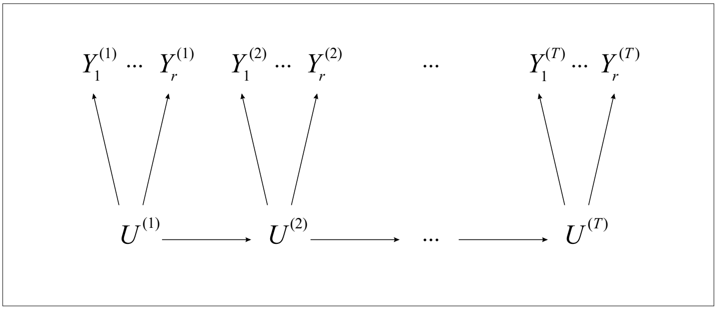
$Y_1^{(1)}$ ... $Y_r^{(1)}$
$Y_1^{(2)}$ ... $Y_r^{(2)}$
...
$Y_1^{(T)}$ ... $Y_r^{(T)}$
$U^{(1)}$
$U^{(2)}$
...
$U^{(T)}$

FIGURE 3.2
Path diagram for the basic LM model for multivariate data

The model assumptions imply that

$$f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U}}(\tilde{\boldsymbol{y}}|\boldsymbol{u}) = \prod_{t=1}^{T} \phi_{\boldsymbol{y}^{(t)}|u^{(t)}}, \tag{3.9}$$

where, in general, we define $\phi_{\boldsymbol{y}|u} = f_{\boldsymbol{Y}^{(t)}|U^{(t)}}(\boldsymbol{y}|u)$; note that, due to assumption of local independence, we have

$$\phi_{\boldsymbol{y}|u} = \prod_{j=1}^{r} \phi_{jy_j|u}. \tag{3.10}$$

In understanding the above expressions, consider that the vector $\tilde{\boldsymbol{y}}$ in (3.9) has a different dimension than the vector $\boldsymbol{y}$ in (3.10). In fact, $\tilde{\boldsymbol{y}}$ is a realization of $\tilde{\boldsymbol{Y}}$ and has subvectors $\boldsymbol{y}^{(t)}$, $t = 1, \ldots, T$, whereas $\boldsymbol{y}$ is a realization of $\boldsymbol{Y}^{(t)}$ and has elements y_j, $j = 1, \ldots, r$. This is in agreement with the general rule adopted throughout the book.

Example 9 — Computation of the conditional distribution of the response variables. *Consider, in the same framework as in Examples 7 and 8, that for every subject we observe $r = 3$ response variables at each occasion. Then, we have*

$$f_{\boldsymbol{Y}^{(t)}|U^{(t)}}(\boldsymbol{y}|u) = \phi_{1y_1|u}\phi_{2y_2|u}\phi_{3y_3|u},$$

and

$$\begin{aligned} f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U}}(\tilde{\boldsymbol{y}}|\boldsymbol{u}) &= \phi_{1y_1^{(1)}|u^{(1)}}\phi_{2y_2^{(1)}|u^{(1)}}\phi_{3y_3^{(1)}|u^{(1)}} \\ &\times \phi_{1y_1^{(2)}|u^{(2)}}\phi_{2y_2^{(2)}|u^{(2)}}\phi_{3y_3^{(2)}|u^{(2)}} \\ &\times \phi_{1y_1^{(3)}|u^{(3)}}\phi_{2y_2^{(3)}|u^{(3)}}\phi_{3y_3^{(3)}|u^{(3)}}. \end{aligned}$$

The manifest probability $f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$ has the same expression as in (3.4), with $\phi_{y^{(t)}|u^{(t)}}$ substituted by $\phi_{\boldsymbol{y}^{(t)}|u^{(t)}}$, as defined in equation (3.10), and may be computed by using the same recursion illustrated in Appendix 1.

3.4 Model identifiability

Identifiability is a necessary requirement of a statistical model, in order to ensure that an estimator of its parameters has desirable asymptotic properties, such as consistency. We are referring, in particular, to the maximum likelihood estimator. Moreover, identifiability is a necessary condition in order to ensure that statistical tests based on asymptotic methods lead to valid results.

In the statistical literature, at least two different definitions of identifiability are known. The first states that a statistical model is *globally identifiable* when it is not possible to find two distinct points of the parameter space which lead to the same distribution of the response variables. A less stringent condition is that of *local identifiability*, which restricts the two points of the parameter space to be one in the neighborhood of the other. For a formal definition of identifiability we refer to Wald (1949), who was one of the first authors to deal with global identifiability in connection with the consistency of the maximum likelihood estimator. See also Rao (1973, Chapter 5). Regarding local identifiability, we refer to McHugh (1956), Goodman (1974), and Rothenberg (1971).

It has to be clear that an LM model may be globally identifiable only up to a switching of the latent states. To clarify this point, in the univariate case consider the distribution of $\tilde{\boldsymbol{Y}}$, $f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$, corresponding to certain values of the parameters $\phi_{y|u}$, π_u, and $\pi_{u|\bar{u}}^{(t)}$. Obviously, if we consider two latent states, say, the first and the second, and we exchange $\phi_{y|1}$ with $\phi_{y|2}$, $y = 0, \dots, c-1$, and accordingly we exchange the initial and transition probabilities, then we obtain the same distribution of $\tilde{\boldsymbol{Y}}$; that is, the value taken by $f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$ does not change for all $\tilde{\boldsymbol{y}}$. This problem, which is common to other latent variable models, such as the LC model, and to finite-mixture models, has been studied in particular in connection with Bayesian inference methods, where it is named the *label switching problem*; see, among others, Stephens (2000) and Jasra et al. (2005).

In order to avoid the label switching problem mentioned above, we may constrain the parameter space by requiring that

$$\phi_{c-1|1} < \dots < \phi_{c-1|k}. \tag{3.11}$$

In this way, the latent states are ordered according to the probability of the last category and then, for instance, the first state includes the subjects with the lowest tendency toward a certain behavior, whereas the last state includes those with the highest tendency.

Given the complexity of the problem, the literature on LM models lacks relevant contributions about conditions which ensure identifiability, apart from the obvious rule that a necessary condition for model identifiability is that the number of parameters must be smaller than the number of possible response configurations of $\tilde{\boldsymbol{y}}$ minus 1. For this reason, we here prefer to propose an empirical approach which is based on comparing the different solutions at convergence of the estimation algorithm starting from different points of the parameter space. This method will be illustrated in Section 3.5.1.3. Methods based on the rank of the observed information matrix will be illustrated in more advanced chapters.

3.5 Maximum likelihood estimation

For an observed sample of n subjects, let $\tilde{\boldsymbol{y}}_i$ denote the (univariate or multivariate) response configuration provided by subject i. Each vector $\tilde{\boldsymbol{y}}_i$ is a realization of $\tilde{\boldsymbol{Y}}$, and then, in the univariate case, it has elements $y_i^{(t)}$ for $t = 1, \dots, T$. In the multivariate case, $\tilde{\boldsymbol{y}}_i$ is made up of the subvectors $\boldsymbol{y}_i^{(t)}$, $t = 1, \dots, T$, which, in turn, have elements $y_{ij}^{(t)}$, $j = 1, \dots, r$.

Assuming independence between the sample units, the log-likelihood of the LM model may be expressed as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}}_i),$$

where $\boldsymbol{\theta}$ is the column vector of all model parameters arranged in a suitable order. We recall that the parameters are the conditional response probabilities $\phi_{y|u}$ (or $\phi_{jy|u}$ in the multivariate case), the initial probabilities π_u, and the transition probabilities $\pi_{u|\bar{u}}^{(t)}$; see Table 3.1 for the list of these parameters in the univariate case.

As usual, it is convenient to rely on the equivalent expression

$$\ell(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{y}}} \log f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}}),$$

where $n_{\tilde{\boldsymbol{y}}}$ denotes the frequency of the response configuration $\tilde{\boldsymbol{y}}$ in the sample. It is already clear that, using this expression for computing $\ell(\boldsymbol{\theta})$, is more efficient since the sum $\sum_{\tilde{\boldsymbol{y}}}$ may be restricted to all response configurations $\tilde{\boldsymbol{y}}$ observed at least once.

We estimate $\boldsymbol{\theta}$ by maximizing the log-likelihood $\ell(\boldsymbol{\theta})$. This may be easily done by the EM algorithm (Baum et al., 1970; Dempster et al., 1977), which has been illustrated for the general case in Section 2.3.

3.5.1 Expectation-Maximization algorithm

In order to illustrate how the EM algorithm may be implemented, we first have to clarify that the complete data are represented by the pairs $(\boldsymbol{u}_i, \tilde{\boldsymbol{y}}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{u}_i = (u_i^{(1)}, \ldots, u_i^{(T)})$ is the sequence of latent states of subject i. In practice, the complete data correspond to the latent state for each subject and time occasion in addition to the observed response configurations. In the following, we first illustrate the EM algorithm for univariate response variables and then for multivariate responses.

3.5.1.1 Univariate formulation

In this case, we have the following expression for the complete data log-likelihood:

$$\begin{aligned} \ell^*(\boldsymbol{\theta}) &= \sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{y=0}^{c-1} a_{uy}^{(t)} \log \phi_{y|u} \\ &+ \sum_{u=1}^{k} b_u^{(1)} \log \pi_u + \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k} b_{\bar{u}u}^{(t)} \log \pi_{u|\bar{u}}^{(t)}, \end{aligned} \tag{3.12}$$

where $a_{uy}^{(t)}$ is the frequency of subjects responding by y at occasion t and belonging to latent state u at the same occasion, $b_u^{(t)}$ is the frequency of subjects in latent state u at occasion t, and $b_{\bar{u}u}^{(t)}$ is the number of transitions from latent state $\bar{u}$ to u at occasion t. In symbols, we have

$$\begin{aligned} a_{uy}^{(t)} &= \sum_{i=1}^{n} I(u_i^{(t)} = u, y_i^{(t)} = y), \\ b_u^{(t)} &= \sum_{i=1}^{n} I(u_i^{(t)} = u), \\ b_{\bar{u}u}^{(t)} &= \sum_{i=1}^{n} I(u_i^{(t-1)} = \bar{u}, u_i^{(t)} = u). \end{aligned}$$

Note that $\ell^*(\boldsymbol{\theta})$ is made up of three components that may be separately maximized. These components involve the parameters $\phi_{y|u}$, the parameters π_u, and the parameters $\pi^{(t)}_{u|\bar{u}}$, respectively; see Bartolucci (2006) for a similar decomposition.

The frequencies $a^{(t)}_{uy}$, $b^{(t)}_{u}$, and $b^{(t)}_{\bar{u}u}$ are obviously unknown. Then, the EM algorithm proceeds by alternating the following two steps until convergence:

- **E-step**: compute the expected value of the above frequencies, given the observed data and the current value of the parameters, so as to obtain the expected value of $\ell^*(\boldsymbol{\theta})$. The expected values of these frequencies are simply

$$\begin{aligned}\hat{a}^{(t)}_{uy} &= \sum_{i=1}^{n} f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}_i)I(y^{(t)}_i = y)\\ &= \sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{y}}} f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}})I(y^{(t)} = y), \qquad (3.13)\end{aligned}$$

$$\hat{b}^{(t)}_{u} = \sum_{i=1}^{n} f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}_i) = \sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{y}}} f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}), \qquad (3.14)$$

$$\begin{aligned}\hat{b}^{(t)}_{\bar{u}u} &= \sum_{i=1}^{n} f_{U^{(t-1)},U^{(t)}|\tilde{\boldsymbol{Y}}}(\bar{u},u|\tilde{\boldsymbol{y}}_i)\\ &= \sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{y}}} f_{U^{(t-1)},U^{(t)}|\tilde{\boldsymbol{Y}}}(\bar{u},u|\tilde{\boldsymbol{y}}). \qquad (3.15)\end{aligned}$$

  In particular, $\hat{a}^{(t)}_{uy}$ must be computed for $t = 1,\ldots,T$, $u = 1,\ldots,k$, and $y = 0,\ldots,c-1$, $\hat{b}^{(t)}_{u}$ must be computed for $t = 1,\ldots,T$ and $u = 1,\ldots,k$, whereas $\hat{b}^{(t)}_{\bar{u}u}$ must be computed for $t = 2,\ldots,T$ and $\bar{u},u = 1,\ldots,k$.

- **M-step**: update the estimate of $\boldsymbol{\theta}$ by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$ obtained as above. Explicit solutions are available for this aim. In particular, we have

  - *Conditional response probabilities*:

$$\phi_{y|u} = \frac{\sum_{t=1}^{T} \hat{a}^{(t)}_{uy}}{\sum_{t=1}^{T} \hat{b}^{(t)}_{u}}, \qquad (3.16)$$

    to be computed for $u = 1,\ldots,k$ and $y = 0,\ldots,c-1$.

  - *Initial probabilities*:

$$\pi_u = \frac{\hat{b}^{(1)}_{u}}{n}, \qquad (3.17)$$

    to be computed for $u = 1,\ldots,k$.

– *Transition probabilities*:

$$\pi_{u|\bar{u}}^{(t)} = \frac{\hat{b}_{\bar{u}u}^{(t)}}{\hat{b}_{\bar{u}}^{(t-1)}}, \tag{3.18}$$

to be computed for $t = 2, \ldots, T$ and $\bar{u}, u = 1, \ldots, k$.

An important point is how to efficiently compute the posterior probabilities $f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}})$ and $f_{U^{(t-1)},U^{(t)}|\tilde{\boldsymbol{Y}}}(\bar{u}, u|\tilde{\boldsymbol{y}})$, involved in (3.13), (3.14), and (3.15). In principle, $\tilde{\boldsymbol{y}}$ may be every response configuration, but in practice it will correspond to one of the observed response configurations $\tilde{\boldsymbol{y}}_i$. First of all consider the probabilities

$$\bar{q}^{(t)}(\bar{u}, \tilde{\boldsymbol{y}}) = f_{Y^{(t+1)},\ldots,Y^{(T)}|U^{(t)}}(y^{(t+1)}, \ldots, y^{(T)}|\bar{u}).$$

For $t = 1, \ldots, T-1$, these probabilities may be computed by the backward recursion

$$\bar{q}^{(t)}(\bar{u}, \tilde{\boldsymbol{y}}) = \sum_{u=1}^{k} \bar{q}^{(t+1)}(u, \tilde{\boldsymbol{y}}) \pi_{u|\bar{u}}^{(t+1)} \phi_{y^{(t+1)}|u}, \quad \bar{u} = 1, \ldots, k, \tag{3.19}$$

initialized with $\bar{q}^{(T)}(\bar{u}, \tilde{\boldsymbol{y}}) = 1$, $\bar{u} = 1, \ldots, k$ (Baum et al., 1970; Levinson et al., 1983; MacDonald and Zucchini, 1997, Sec. 2.2). Then, for $t = 1, \ldots, T$, we have

$$f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}) = \frac{q^{(t)}(u, \tilde{\boldsymbol{y}})\bar{q}^{(t)}(u, \tilde{\boldsymbol{y}})}{f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})}, \quad u = 1, \ldots, k, \tag{3.20}$$

whereas, for $t = 2, \ldots, T$ and $\bar{u}, u = 1, \ldots, k$, we have

$$f_{U^{(t-1)},U^{(t)}|\tilde{\boldsymbol{Y}}}(\bar{u}, u|\tilde{\boldsymbol{y}}) = \frac{q^{(t-1)}(\bar{u}, \tilde{\boldsymbol{y}})\pi_{u|\bar{u}}^{(t)}\phi_{y^{(t)}|u}\bar{q}^{(t)}(u, \tilde{\boldsymbol{y}})}{f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})}. \tag{3.21}$$

Even in this case, the recursions can be efficiently implemented by the matrix notation (Bartolucci, 2006; Bartolucci et al., 2007); see Appendix 1. See also Bartolucci and Besag (2002) for alternative recursions to compute the posterior probabilities of interest.

Along the same lines as in Shi et al. (2005), it is possible to prove that the EM algorithm converges to a local maximum of $\ell(\boldsymbol{\theta})$. However, as typically happens for discrete latent variable models, this function may be multimodal and then the convergence is not ensured to be at the global maximum of this function. In this regard, it is crucial to choose the starting points for the EM algorithm in a sensible way. This issue will be discussed in detail in Section 3.5.1.3.

Another important point is how to check for convergence. The two most common criteria used for this aim are based on the difference in terms of log-likelihood of two consecutive steps and the distance between the corresponding parameter vectors. More precisely, let $\boldsymbol{\theta}^{(s)}$ denote the parameter estimate obtained at the end of the s-th M-step. Then, according to the first criterion, the algorithm is stopped when

$$\ell(\boldsymbol{\theta}^{(s)}) - \ell(\boldsymbol{\theta}^{(s-1)}) \geq \varepsilon_1 > 0, \tag{3.22}$$

where ε_1 is a suitable tolerance level; in our applications we use $\varepsilon_1 = 10^{-6}$. According to the second criterion, the algorithm is stopped when

$$\max_h |\theta_h^{(s)} - \theta_h^{(s-1)}| \geq \varepsilon_2 > 0, \tag{3.23}$$

that is, when the maximum of elementwise distance between $\boldsymbol{\theta}^{(s)}$ and $\boldsymbol{\theta}^{(s-1)}$ is less than the tolerance ε_2, which may be chosen as above.

Obviously, an accurate check of convergence needs to rely on both conditions above. In fact, as frequently happens for latent variable models, the log-likelihood may present a rather flat region around a local maximum, implying that condition (3.22) is satisfied even when $\boldsymbol{\theta}^{(s)}$ and $\boldsymbol{\theta}^{(s-1)}$ are not so close.

Finally, the use of very small values for ε_1 and ε_2 takes into account the slowness to converge of the EM algorithm in comparison with other optimization algorithms which, however, are usually unstable. For this reason, more sophisticated rules are sometimes used to check the convergence of the EM algorithm, such as that based on an estimate of the distance of $\ell(\boldsymbol{\theta}^{(s)})$ from the supremum of the log-likelihood function; see, for instance, Böhning et al. (1994).

3.5.1.2 Multivariate formulation

With multivariate responses, we have the following expression for the complete data log-likelihood

$$\begin{aligned} \ell^*(\boldsymbol{\theta}) &= \sum_{j=1}^{r}\sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{y=0}^{c_j-1} a_{juy}^{(t)} \log \phi_{jy|u} \\ &+ \sum_{u=1}^{k} b_u^{(1)} \log \pi_u + \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k} b_{\bar{u}u}^{(t)} \log \pi_{u|\bar{u}}^{(t)}, \end{aligned}$$

which is different from (3.12) in the first sum, which involves the frequencies $a_{juy}^{(t)}$. In particular, $a_{juy}^{(t)}$ corresponds to the number of subjects that, at occasion t, are in latent state u and have outcome y for the j-th response variable, that is,

$$a_{juy}^{(t)} = \sum_{i=1}^{n} I(u_i^{(t)} = u, y_{ij}^{(t)} = y). \tag{3.24}$$

The other components of $\ell^*(\boldsymbol{\theta})$ remain unchanged.

The EM algorithm has the same structure as the one outlined in Section 3.5.1.1, and the same criteria for checking the convergence may be adopted. The only difference is that, at the E-step, we need to compute the expected frequencies

$$\hat{a}_{juy}^{(t)} = \sum_{i=1}^{n} f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}_i) I(y_{ij}^{(t)} = y) = \sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{y}}} f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}) I(y_j^{(t)} = y),$$

for $j = 1, \ldots, r$, $t = 1, \ldots, T$, $u = 1, \ldots, k$, and $y = 0, \ldots, c_j - 1$, instead of those in (3.13). In practice, these expected frequencies, together with the expected frequencies in (3.14) and (3.15), are computed by the same recursions as in Section 3.5.1.1; see also Appendix 1.

With the M-step, the only difference is that the conditional response probabilities are updated as follows:

$$\phi_{jy|u} = \frac{\sum_{t=1}^{T} \hat{a}_{juy}^{(t)}}{\sum_{t=1}^{T} \hat{b}_u^{(t)}}, \tag{3.25}$$

for $j = 1, \ldots, r$, $u = 1, \ldots, k$, and $y = 0, \ldots, c_j - 1$, rather than through (3.16).

3.5.1.3 Initialization of the algorithm and model identifiability

Regarding the initialization of the EM algorithm, we suggest adopting a multistart strategy which combines a deterministic rule with a random starting rule; these rules are described in detail below. In this way, we attempt to prevent problems due to the multimodality of the likelihood function. In fact, the random rule allows us to adequately explore the parameter space, when its application is repeated a suitable number of times. Therefore, once an estimate is obtained starting with the deterministic rule, denoted by $\hat{\boldsymbol{\theta}}_0$, we suggest performing the algorithm again starting from a suitable number R of randomly chosen points of the parameters space, obtaining the estimates $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_R$. Then, we compare $\ell(\hat{\boldsymbol{\theta}}_0)$ with the maximum of $\ell(\hat{\boldsymbol{\theta}}_1), \ldots, \ell(\hat{\boldsymbol{\theta}}_R)$. Provided that R is large enough, if $\ell(\hat{\boldsymbol{\theta}}_0)$ is not smaller than this maximum (up to a negligible tolerance level), then we can be confident that the solution based on the deterministic starting rule corresponds to the global maximum of $\ell(\boldsymbol{\theta})$.

Otherwise, this rule needs to be somehow improved. In any case, we take as estimate of the parameters, denoted by $\hat{\boldsymbol{\theta}}$, the one corresponding to the highest log-likelihood among $\hat{\boldsymbol{\theta}}_0,\dots,\hat{\boldsymbol{\theta}}_R$. A similar strategy was proposed by Berchtold (2004) in a related field.

In the case of univariate responses, the deterministic rule we suggest using for initializing the EM algorithm consists of computing the global logits for the observed distribution of the response variables. In practice, this amounts to computing

$$\eta_y = \log\frac{\sum_{i=1}^n\sum_{t=1}^T I(y_i^{(t)}\geq y)}{\sum_{i=1}^n\sum_{t=1}^T I(y_i^{(t)}< y)},\quad y=1,\dots,c-1.$$

Moreover, once a grid of points $\nu_1,\dots,\nu_k$ is chosen (we suggest a grid of equally spaced points between $-k$ and k), the conditional response probabilities are initialized as follows:

$$\phi_{y|u}=\phi^*_{y|u}-\phi^*_{y+1|u},\quad y=0,\dots,c-1, \tag{3.26}$$

where

$$\phi^*_{y|u}=\begin{cases}1 & y=0,\\ \frac{\exp(\eta_y+\nu_u)}{1+\exp(\eta_y+\nu_u)} & y=1,\dots,c-1,\\ 0 & y=c,\end{cases}$$

for $u=1,\dots,k$. Note that $\phi^*_{y|u}$ are the values of the corresponding survival function, that is, $\phi^*_{y|u}=p(Y^{(t)}\geq y|U^{(t)}=u)$. This rule guarantees that the conditional probabilities $\phi_{y|u}$ sum up to 1 and the resulting distribution of the response variables is statistically increasing with u. In the multivariate case, we suggest applying this rule separately for each response variable, so as to initialize the conditional response probabilities $\phi_{jy|u}$, $j=1,\dots,r$, $u=1,\dots,k$, $y=0,\dots,c_j-1$.

Finally, regardless of the nature of the response variables, we suggest using the following starting values for the initial probabilities:

$$\pi_u=\frac{1}{k},\quad u=1,\dots,k, \tag{3.27}$$

whereas for the transition probabilities, we suggest using

$$\pi^{(t)}_{u|\bar{u}}=\frac{1}{h+k}\begin{cases}h+1, & u=\bar{u},\\ 1, & u\neq\bar{u},\end{cases} \tag{3.28}$$

for $t=2,\dots,T$, where h is a suitable constant (we use $h=9$ in our applications).

The random starting rule that we propose is based on suitably normalized random numbers drawn from a uniform distribution between

0 and 1. In particular, in the univariate case we first draw each $\phi_{y|u}$, $u = 1, \dots, k$, $y = 0, \dots, c-1$, from this distribution, and then we normalize these probabilities so that the constraint $\sum_{y=0}^{c-1} \phi_{y|u} = 1$ is satisfied for $u = 1, \dots, k$. The same procedure is applied in the multivariate case to choose the starting values of $\phi_{jy|u}$, $j = 1, \dots, r$, $u = 1, \dots, k$, $y = 0, \dots, c-1$. In a similar way, we suggest choosing each initial probability π_u, $u = 1, \dots, k$, as a random number drawn from a uniform distribution between 0 and 1 which is suitably normalized. The same may be applied, for $t = 2, \dots, T$ and $\bar{u} = 1, \dots, k$, to draw the transition probabilities $\pi_{u|\bar{u}}^{(t)}$, $u = 1, \dots, k$, which must sum up to 1.

As mentioned in Section 3.4, a fundamental issue in applying a statistical model is to check if this model is identifiable. Since no general rules are available for the basic LM model, we propose here a simple empirical approach which is based on checking the distance, in the sense of (3.23), between the parameter estimates obtained starting from different points of the parameter space, as described above. In particular, if among $\hat{\boldsymbol{\theta}}_0, \dots, \hat{\boldsymbol{\theta}}_R$ there exist at least two different estimates leading to a value of the log-likelihood close to $\ell(\hat{\boldsymbol{\theta}})$, then we can assess that the model is not globally identifiable. Otherwise, provided that R is large enough, we can be confident about the global identifiability of the model. It has to be clear that, in performing this comparison, the elements of the estimated parameter vector must be suitably ordered on the basis of a constraint of type (3.11). For a similar strategy to check local identifiability, see Forcina (2008).

3.5.2 Alternative algorithms for maximum likelihood estimation

The EM algorithm is not the only possible algorithm to maximize the log-likelihood of an LM model. In the hidden Markov (HM) literature, different algorithms have been applied; these algorithms may be applied in our context as well. For a related discussion see Zucchini and MacDonald (2009, Chapter 3).

The most natural solution is direct maximization through the Newton-Raphson (NR) algorithm, which, for simpler models, is a much faster maximization procedure with respect to the EM algorithm. The ingredients needed for implementing the NR algorithm are the first derivative vector and the second derivative matrix of the log-likelihood. The latter is substituted by the expected information matrix in the Fisher-scoring algorithm. See Section 7.4.1 for details.

Regarding direct maximization via the NR algorithm, or via the Fisher-scoring algorithm, of the log-likelihood of the basic LM model, Lystig and Hughes (2002) and Cappé and Moulines (2005) developed recursive procedures for calculating the derivatives of the log-likelihood for HM models; see also Bartolucci (2006) and Turner (2008).

Other alternatives to the EM algorithm for the maximization of the likelihood of LM models include the proposal of Zucchini and Guttorp (1991), who made use of a general numerical algorithm. See also Collings and Rydén (1998) and Altman and Petkau (2005).

In this book, however, we prefer to rely on the EM algorithm because the other algorithms mentioned above are more difficult to implement and less stable. In any case, it is worth noting that this algorithm can be combined with an iterative algorithm such as the NR, so as to have a faster convergence to a local maximum of the model log-likelihood, while retaining an acceptable level of stability. At the beginning, the EM algorithm is run until the log-likelihood considerably increases from step to step. Then, the NR algorithm is run starting from the EM solution until the final convergence.

3.5.3 Standard errors

Given the nature of the parameter space of the basic LM model, which is bounded because all parameters correspond to (marginal or conditional) probabilities, the preferred method to obtain standard errors for the maximum likelihood estimate is the parametric bootstrap. This point has already been briefly discussed in Section 2.4. Moreover, we provide a detailed illustration of this method in Section 7.4.2.

3.6 Selection of the number of latent states

As for the basic LC model, a fundamental point in applying the LM model is the choice of the number of latent states, denoted by k. With univariate responses, this number is sometimes fixed equal to c, that is, the number of response categories. This choice is typically adopted when the LM model is seen as an extension of an MC model which accounts for measurement errors. In such a case, the latent states are seen as the "true" states of the chain and the observed states are seen as the "noisy" states. Consequently, the conditional response probabilities $\phi_{y|u}$ are interpreted as transition probabilities from the "true" to the "noisy" states.

When the number of latent states cannot be *a priori* defined, the same selection criteria which are defined in Section 2.5.4 may be adopted. In this section, we recall that, even if in principle a criterion based on the likelihood ratio test between nested models may be adopted, information criteria are used in most applications. In particular, the two most common criteria are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). These criteria are based on indices which have the same expression as in Section 2.5.4, that is,

$$\begin{aligned} AIC &= -2\hat{\ell} + 2\#\text{par}, \\ BIC &= -2\hat{\ell} + \log(n)\#\text{par}, \end{aligned}$$

where $\hat{\ell}$ denotes the maximum of the log-likelihood of the LM model of interest and #par denotes the number of free parameters.

Even if these criteria are widely used, their performances have not been studied enough in detail in connection with LM models. On the other hand, their theoretical properties have been studied with reference to HM models; see Celeux and Durand (2006) and Boucheron and Gassiat (2005). However, we have to recall that the context is different since HM models are used for time series. Given this lack in the literature about LM models, in Section 7.6 we provide an illustration of the performance of AIC and BIC which is based on a simulation study.

Finally, it is worth noting that, depending of the specific application of interest, different selection criteria may be used which take the quality of the classification into account. For instance, Bartolucci et al. (2009) proposed to rely on the index

$$S = \frac{\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{y}}} \sum_{t=1}^{T} [\hat{f}^{*(t)}(\tilde{\boldsymbol{y}}) - 1/k]}{(1 - 1/k)nT},$$

where $f^{*(t)}(\tilde{\boldsymbol{y}})$ is the maximum, with respect to u, of the posterior probabilities $f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}})$. Index S is always between 0 and 1, with 1 corresponding to the situation of absence of uncertainty in the classification, since one of such posterior probabilities is equal to 1 for every $\tilde{\boldsymbol{y}}$ and t, with all the other probabilities equal to 0. Criteria that take into account the quality of the classification can also be based on entropy measures; with reference to the LC model, see Celeux and Soromenho (1996).

3.7 Applications

In the following, we illustrate the approach of this chapter by analyzing the two datasets already considered in Chapter 2.

3.7.1 Marijuana consumption dataset

We here show the results of the application of the EM algorithm for estimating the basic LM model using the marijuana consumption dataset illustrated in Section 1.4.1.

In applying the basic LM model, here denoted by M_5, it is natural to choose $k = 3$ latent states that may be seen as "true" levels of drug consumption. With this number of latent states, we applied the EM algorithm starting with the deterministic rule described in Section 3.5.1.3; the algorithm reaches a maximum log-likelihood equal to $\hat{\ell} = -646.89$, after 631 iterations. With 32 parameters, we then have AIC and BIC indices equal to 1357.79 and 1468.77, respectively. The adopted criterion for stopping the algorithm is that the difference between two consecutive steps, in terms of log-likelihood and distance between parameter vectors, is less than 10^{-6}. More precisely, we stopped the EM algorithm when both (3.22) and (3.23) were satisfied with $\varepsilon_1 = \varepsilon_2 = 10^{-6}$. Anyway, the algorithm converged in a few seconds, given the availability of explicit solutions which are used at the M-step. The plot of the log-likelihood, represented against the iteration number, is represented in Figure 3.3.

From the figure, we observe that the EM algorithm very rapidly increases the log-likelihood, but it becomes very slow when it is close to

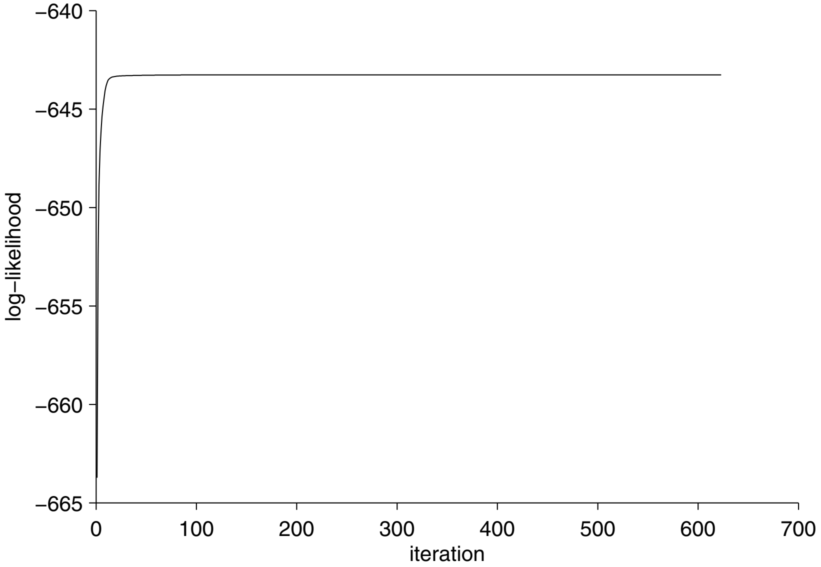

FIGURE 3.3
Log-likelihood against the iteration number

convergence. This is confirmed by the fact that, applying the same stopping as above but with $\varepsilon_1 = \varepsilon_2 = 10^{-10}$, the algorithm needs many more iterations, that is, 11863, and reaches a maximum log-likelihood which is only larger by 0.000226 than the above one. Moreover, the distance between the two estimates, in the sense of (3.23), is equal to 0.006133.

In order to illustrate the problem of local maxima, we fitted the same model starting from other $R = 25$ different points which were randomly chosen, as described in Section 3.5.1.3, using a stopping rule based on (3.22) and (3.23), with $\varepsilon_1 = \varepsilon_2 = 10^{-10}$. The results, ordered according to the log-likelihood at convergence, are reported in Table 3.2, together with the maximum value reached starting with the deterministic rule; we also show the difference with respect to the best solution in terms

TABLE 3.2
Maximum log-likelihood obtained starting with the deterministic and a series of random initializations, together with the difference with respect to the best solution also in terms of parameter estimates

		Difference wrt best	
	Log-likelihood	Log-likelihood	Parameters
deterministic	−646.893797	0.000000	0.000000
random	−646.893797	–	–
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893797	0.000000	0.000000
	−646.893896	−0.000099	0.006134
	−646.893924	−0.000127	0.006134
	−646.893924	−0.000127	0.006134
	−646.893924	−0.000127	0.006134
	−646.893924	−0.000127	0.006134
	−646.893924	−0.000127	0.006134
	−646.893924	−0.000127	0.006134
	−646.901838	−0.008040	0.816190
	−646.901838	−0.008040	0.816190

of log-likelihood and distance between the parameter estimates. These estimates are ordered so that constraint (3.11) is satisfied.

The results in Table 3.2 show evidence of more local maxima, which, however, are very close. Moreover, the deterministic initialization leads to the highest log-likelihood. We take the corresponding solution as the maximum likelihood estimate of the parameters. Finally, we did not find different parameter vectors which correspond to a log-likelihood equal to the maximum value. This leads us to be confident about the identifiability of the basic LM model, as clarified in Section 3.5.1.3.

For the data at hand, the estimates of the conditional response probabilities obtained as above are reported in Table 3.3, whereas the estimates of the initial and transition probabilities are reported in Tables 3.4 and 3.5, respectively. The corresponding standard errors based on the bootstrap method are reported in Section 7.4.2, where this method is illustrated in detail.

TABLE 3.3
Estimates of the conditional response probabilities $\phi_{y|u}$ under model M_5

	$\hat{\phi}_{y\|u}$		
u	$y = 0$	$y = 1$	$y = 2$
1	0.9959	0.0000	0.0041
2	0.3053	0.6870	0.0077
3	0.0116	0.0834	0.9050

TABLE 3.4
Estimates of the initial probabilities π_u under model M_5

u	$\hat{\pi}_u$
1	0.8978
2	0.0837
3	0.0185

The results in Table 3.3 allow us to interpret the latent states in terms of tendency to marijuana consumption. In particular, we observe that the first state corresponds to the lowest tendency of marijuana consumption, the second state to an intermediate tendency, and the third to the highest tendency. This is because, when u goes from 1 to 3, the conditional probability of the first category ($y = 0$) tends to decrease, whereas that of the last category ($y = 2$) tends to increase. Even if a formal testing procedure would be necessary, we also conclude that there is evidence of measurement errors, especially for subjects in the second

TABLE 3.5
Estimates of the transition probabilities $\pi_{u|\bar{u}}^{(t)}$ under model M_5

		$\hat{\pi}_{u\|\bar{u}}^{(t)}$		
t	$\bar{u}$	$u=1$	$u=2$	$u=3$
2	1	0.8305	0.1545	0.0150
	2	0.3196	0.2275	0.4529
	3	0.0560	0.0000	0.9440
3	1	0.8099	0.1902	0.0000
	2	0.0575	0.4815	0.4610
	3	0.0000	0.1471	0.8529
4	1	0.9074	0.0644	0.0281
	2	0.0589	0.7176	0.2235
	3	0.0000	0.1857	0.8143
5	1	0.7886	0.1629	0.0484
	2	0.0998	0.8202	0.0800
	3	0.0198	0.0356	0.9446

and third states. For instance, subjects in the third latent state have probability 0.9050 to declare a consumption corresponding to the third category. In absence of measurement errors we would have a true value of this probability equal to 1.

From the estimates in Tables 3.4 and 3.5, we observe that most of subjects belong to the first latent state at the beginning of the study. Then, given that we have more than one transition matrix, the interpretation of these elements is not so easy.

In order to simplify the interpretation of the results, we can consider the marginal distribution of the latent states which may be derived by using standard rules for Markov chains. In particular, for each wave we obtain the marginal distribution given in Table 3.6.

TABLE 3.6
Estimated marginal distribution of the latent states for each time occasion under model M_5

	$\hat{f}_{U^{(t)}}(u)$		
t	$u=1$	$u=2$	$u=3$
1	0.8978	0.0837	0.0185
2	0.7734	0.1578	0.0688
3	0.6354	0.2332	0.1314
4	0.5903	0.2327	0.1770
5	0.4923	0.2933	0.2144

It is clear from the results in Table 3.6 that the tendency to use marijuana increases with age. In fact, the probability of the first state systematically decreases, whereas that of the third state increases. This behavior is clearly confirmed by the plot in Figure 3.4, which represents these distributions.

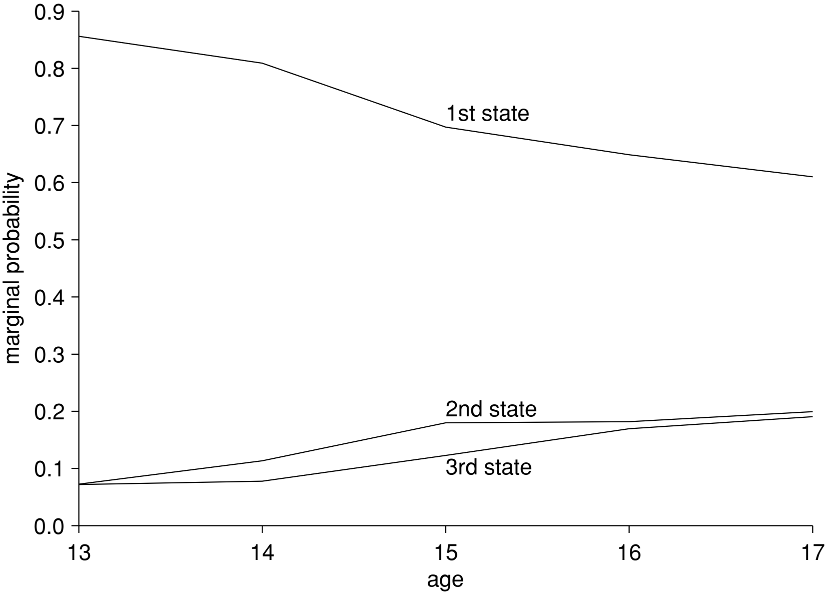

FIGURE 3.4
Representation of the estimates of the probabilities $f_{U^{(t)}}(u)$ versus t for each latent state u under model M_5

A final comment concerns the comparison between the model here considered, M_5, and models M_1 and M_3 considered in Chapter 2. In particular, model M_1 which follows an LC formulation (see Section 2.6.1) has the same number of parameters as M_5, but a smaller log-likelihood and a higher AIC and BIC. We can then conclude that, for these data, a basic LM formulation is better than a basic LC formulation in terms of fit. Moreover, in comparison with model M_1, model M_5 provides a more realistic interpretation of the phenomenon under investigation.

Compared with model M_3, which is based on an MC formulation (see Section 2.8.1), we observe that model M_5 has a smaller AIC index, but a higher BIC index. However, we delay the formulation of a more sophisticated LM model for these data to Chapter 4.

3.7.2 Criminal conviction history dataset

In analyzing this dataset, we considered all ten typologies of crime for each of the six age bands; see Section 1.4.2. Therefore, we applied the basic LM model in its multivariate version described in Section 3.3, with $r = 10$ and $T = 6$. Since all response variables are binary, we have $c_j = 2$ for $j = 1, \ldots, r$.

The first step of the analysis was aimed at the choice of a suitable number of latent states, k. Therefore, we fitted the multivariate basic LM model for increasing values of k until the BIC index decreased with respect to the previous value. We obtained the results in Table 3.7.

TABLE 3.7
Maximum log-likelihood, number of parameters, and AIC and BIC indices for a number of latent states between 1 and 7

k	$\hat{\ell}$	#par	AIC	BIC
1	−145466.50	10	290953.00	291041.71
2	−114894.63	31	229851.25	230126.27
3	−112998.80	62	226121.59	226671.63
4	−111633.85	103	223473.71	224387.48
5	−111261.97	154	222831.95	224198.17
6	−110676.02	215	221782.03	223689.42
7	−110380.55	286	221333.10	223870.37

On the basis of these results we chose $k = 6$ latent states. The resulting multivariate basic LM model is denoted by M_6 in the following. Note that AIC leads to choosing a larger number of classes; however, as usual, we prefer to rely on BIC for this choice. The parameter estimates under model M_6 are reported in Tables 3.8, 3.9, and 3.10. In reporting these estimates, we order the latent states according to the probability of the fifth type of crime, which is the most frequently observed.

On the basis of the results in Table 3.8, we conclude that the first two latent states, and in particular the first one, correspond to subjects with null or very low tendency to commit crimes. On the other hand, with only one exception, the last latent state corresponds to subjects with the highest tendency to commit each type of crime. The other latent states correspond to subjects with an intermediate tendency to commit crimes and may be characterized in terms of specialization toward specify types of crime. In particular, subjects in the third latent state are specialized in crimes of type 1 (Violence against the person), 2 (Sexual offenses), and 8 (Drug offenses). Subjects in the fourth latent state are specialized in crimes of type 6 (Fraud and forgery) and 9 (Motoring offenses). Finally,

TABLE 3.8
Estimates of the conditional response probabilities $\phi_{j1|u}$ under model M_6

	$\hat{\phi}_{j1\|u}$					
j	$u=1$	$u=2$	$u=3$	$u=4$	$u=5$	$u=6$
1	0.0015	0.0000	0.2315	0.0443	0.0534	0.3770
2	0.0006	0.0022	0.0246	0.0111	0.0121	0.0382
3	0.0000	0.0206	0.0209	0.0573	0.4862	0.5526
4	0.0000	0.0001	0.0087	0.0039	0.0164	0.0660
5	0.0034	0.1148	0.1380	0.6461	0.6949	0.8319
6	0.0007	0.0000	0.0160	0.3486	0.0137	0.2459
7	0.0008	0.0127	0.1335	0.0252	0.1375	0.3263
8	0.0007	0.0000	0.1106	0.0338	0.0000	0.1906
9	0.0000	0.0000	0.0079	0.0190	0.0000	0.0913
10	0.0004	0.0000	0.1078	0.1023	0.1909	0.4760

TABLE 3.9
Estimates of the initial probabilities π_u under model M_6

u	$\hat{\pi}_u$
1	0.7831
2	0.1878
3	0.0005
4	0.0003
5	0.0277
6	0.0006

subjects in the fifth latent states are specialized in crimes of type 3 (Burglary), 4 (Robbery), 7 (Criminal damage), and 10 (Other offenses).

On the basis of the estimates in Table 3.9, we conclude that, at the beginning of the period of observation corresponding to the age band 10–15, the 97% of subjects are in the first two latent states which correspond to a very low tendency to commit crimes. The other states have a negligible initial probability, with the exception of the fifth, corresponding to a moderate tendency toward Burglary, Robbery, Criminal Damage, and Other offenses.

Finally, from the estimates in Table 3.10, we observe how the phenomenon under investigation evolves and, in particular, how aging affects the tendency to commit crimes.

The first consideration is that there is a high persistence in the first latent state, meaning that subjects having the lowest tendency to commit crimes very rarely move to states with a higher criminal tendency. A

TABLE 3.10
Estimates of the transition probabilities $\pi^{(t)}_{u|\bar{u}}$ under model M_6

		$\hat{\pi}^{(t)}_{u\|\bar{u}}$					
t	$\bar{u}$	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$	$u = 6$
2	1	0.8718	0.1034	0.0164	0.0083	0.0000	0.0000
	2	0.1904	0.5933	0.0829	0.0300	0.0770	0.0266
	3	0.2631	0.0000	0.4597	0.2772	0.0000	0.0000
	4	0.4886	0.0000	0.5114	0.0000	0.0000	0.0000
	5	0.0000	0.2816	0.1013	0.0000	0.2790	0.3381
	6	0.0960	0.0000	0.0000	0.0000	0.0000	0.9040
3	1	0.9541	0.0295	0.0119	0.0045	0.0000	0.0000
	2	0.5301	0.3602	0.0666	0.0376	0.0000	0.0056
	3	0.4869	0.1655	0.3213	0.0110	0.0000	0.0153
	4	0.4819	0.2372	0.0970	0.1630	0.0000	0.0210
	5	0.0000	0.5041	0.1749	0.0743	0.1440	0.1027
	6	0.0000	0.1235	0.2390	0.0044	0.1247	0.5084
4	1	0.9826	0.0114	0.0028	0.0032	0.0000	0.0000
	2	0.5168	0.3868	0.0460	0.0474	0.0009	0.0021
	3	0.5904	0.0049	0.3704	0.0233	0.0050	0.0059
	4	0.5459	0.1366	0.0616	0.2278	0.0203	0.0078
	5	0.0000	0.2584	0.2455	0.1090	0.3267	0.0605
	6	0.0000	0.1966	0.2625	0.0724	0.0092	0.4593
5	1	0.9968	0.0000	0.0013	0.0018	0.0001	0.0001
	2	0.4704	0.4351	0.0592	0.0354	0.0000	0.0000
	3	0.4001	0.1450	0.4312	0.0083	0.0030	0.0124
	4	0.4101	0.3102	0.0000	0.2797	0.0000	0.0000
	5	0.0000	0.6286	0.0758	0.0000	0.2720	0.0236
	6	0.0000	0.0795	0.3384	0.1239	0.0674	0.3908
6	1	0.9988	0.0000	0.0000	0.0011	0.0000	0.0001
	2	0.8527	0.0000	0.0808	0.0665	0.0000	0.0000
	3	0.5947	0.0000	0.3825	0.0205	0.0000	0.0022
	4	0.6126	0.0942	0.0405	0.2496	0.0000	0.0031
	5	0.2227	0.0000	0.4003	0.0000	0.3068	0.0702
	6	0.0097	0.2389	0.2001	0.1588	0.0000	0.3925

certain tendency, but much less evident, is observed with reference to the last latent state, to which subjects with the highest tendency to commit crimes belong. Another interesting observation is that, starting from the second age band, subjects in the second to the fourth latent state have a large probability to move to the first state, revealing the tendency to considerably improve in their behavior (see transition matrices for $t \geq 3$). On the other hand, this pattern is not observed with reference to the transition from the first to the second age band ($t = 2$).

As for the example in Section 3.7.1, the evolution of the phenomenon may be further studied through the analysis of the estimated marginal distribution of $U^{(t)}$ obtained, for each t, on the basis of the initial and transition probabilities indicated above. These distributions are reported in Table 3.11.

TABLE 3.11
Estimated marginal distribution of the latent states for each time occasion under model M_6

	$\hat{f}_{U^{(t)}}(u)$					
t	$u=1$	$u=2$	$u=3$	$u=4$	$u=5$	$u=6$
1	0.7831	0.1878	0.0005	0.0003	0.0277	0.0006
2	0.7188	0.2002	0.0316	0.0123	0.0222	0.0149
3	0.8132	0.1144	0.0407	0.0148	0.0051	0.0117
4	0.8904	0.0594	0.0278	0.0138	0.0024	0.0063
5	0.9322	0.0361	0.0189	0.0086	0.0012	0.0029
6	0.9787	0.0015	0.0116	0.0064	0.0004	0.0014

We observe that the probability of the first latent state decreases from the first to the second age band, and then increases to almost 0.98 at the end of the period of observation. Consequently, an opposite trend is observed for the probabilities of the other latent states which become negligible at the end of the period. Overall, we conclude that, after an initial period in which the subjects tend to worsen in terms of social behavior, they tend to improve so that at age 40, only around the 2% of them show some tendency to commit crimes.

Appendix 1: Efficient implementation of recursions

Manifest distribution of the response variables

In order to efficiently implement the recursion based on (3.6) and (3.7), which is finalized to compute the manifest probability $f_{\boldsymbol{Y}}(\tilde{\boldsymbol{y}})$, let $\boldsymbol{q}^{(t)}(\tilde{\boldsymbol{y}})$ denote the column vector with elements $q^{(t)}(u, \tilde{\boldsymbol{y}})$, $u = 1, \dots, k$, and in a similar way define the initial probability vector $\boldsymbol{\pi}$ with elements π_u and the conditional probability vector $\boldsymbol{m}_y$ with elements $\phi_{y|u}$. Also let $\boldsymbol{\Pi}^{(t)}$ denote the transition probability matrix with elements $\pi^{(t)}_{u|\bar{u}}$,

$\bar{u}, u = 1, \ldots, k$, with $\bar{u}$ running by row and u by column. Then, we have

$$\boldsymbol{q}^{(t)}(\tilde{\boldsymbol{y}}) = \begin{cases} \text{diag}(\boldsymbol{m}_{y^{(1)}})\boldsymbol{\pi}, & t = 1, \\ \text{diag}(\boldsymbol{m}_{y^{(t)}})(\boldsymbol{\Pi}^{(t)})'\boldsymbol{q}^{(t-1)}(\tilde{\boldsymbol{y}}), & t = 2, \ldots, T. \end{cases} \tag{3.29}$$

At the end of this recursion we obtain $f_{\boldsymbol{Y}}(\tilde{\boldsymbol{y}})$ as $\boldsymbol{q}^{(T)}(\tilde{\boldsymbol{y}})'\boldsymbol{1}_k$, where $\boldsymbol{1}_k$ denotes a column vector of ones of dimension k. In implementing this recursion, attention must be paid to the case of large values of T because, as t increases, the probabilities $q^{(t)}(u, \tilde{\boldsymbol{y}})$ could become negligible; see Scott (2002) for remedial measures.

In the multivariate case, the same recursion as in (3.29) may be used, with $\boldsymbol{m}_y$ substituted by the vector $\boldsymbol{m}_{\boldsymbol{y}}$ with elements $\phi_{\boldsymbol{y}|u}$, $u = 1, \ldots, k$, as defined in (3.10). The same holds for the recursions illustrated below.

Marginal distribution of the latent variables

In order to implement the recursion based on (3.8) using the matrix notation, let $\boldsymbol{q}^{(t)}$ be defined as the vector with elements $q^{(t)}(u)$, $u = 1, \ldots, k$. Then, we have

$$\boldsymbol{q}^{(t)} = \begin{cases} \boldsymbol{\pi}, & t = 1, \\ (\boldsymbol{\Pi}^{(t)})'\boldsymbol{q}^{(t-1)}, & t = 2, \ldots, T. \end{cases}$$

Posterior distribution of the latent variables

Let $\bar{\boldsymbol{q}}^{(t)}(\tilde{\boldsymbol{y}})$ be the column vector with elements $\bar{q}^{(t)}(\bar{u}, \tilde{\boldsymbol{y}})$, $\bar{u} = 1, \ldots, k$. This vector may be computed by the backward recursion

$$\bar{\boldsymbol{q}}^{(t)}(\tilde{\boldsymbol{y}}) = \begin{cases} \boldsymbol{1}_k, & t = T, \\ \boldsymbol{\Pi}^{(t+1)}\text{diag}(\boldsymbol{m}_{y^{(t+1)}})\bar{\boldsymbol{q}}^{(t+1)}(\tilde{\boldsymbol{y}}), & t = T-1, \ldots, 1. \end{cases}$$

Then, the k-dimensional column vector $\boldsymbol{f}^{(t)}(\tilde{\boldsymbol{y}})$ with elements $f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}})$, $u = 1, \ldots, k$, defined in (3.20) is obtained as

$$\boldsymbol{f}^{(t)}(\tilde{\boldsymbol{y}}) = \frac{1}{f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})}\text{diag}[\boldsymbol{q}^{(t)}(\tilde{\boldsymbol{y}})]\bar{\boldsymbol{q}}^{(t)}(\tilde{\boldsymbol{y}}), \quad t = 1, \ldots, T.$$

Moreover, the $k \times k$ matrix $\boldsymbol{F}^{(t)}(\tilde{\boldsymbol{y}})$, with elements $f_{U^{(t-1)},U^{(t)}|\tilde{\boldsymbol{Y}}}(\bar{u}, u|\tilde{\boldsymbol{y}})$ arranged by letting $\bar{u}$ run by row and u by column, is obtained as

$$\boldsymbol{F}^{(t)}(\tilde{\boldsymbol{y}}) = \frac{1}{f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})}\text{diag}[\boldsymbol{q}^{(t-1)}(\tilde{\boldsymbol{y}})]\boldsymbol{\Pi}^{(t)}\text{diag}[\boldsymbol{m}_{y^{(t)}}]\text{diag}[\bar{\boldsymbol{q}}^{(t)}(\tilde{\boldsymbol{y}})],$$

for $t = 2, \ldots, T$.

4

Constrained latent Markov models

4.1 Introduction

This chapter deals with versions of the latent Markov (LM) model which are formulated by incorporating constraints corresponding to certain hypotheses of interest. These constraints may concern the measurement model or the latent model. In the first case, they are posed on the distribution of the response variables given the latent process, whereas in the second case, they are posed on the distribution of the latent process. Obviously, depending on the application of interest, an LM model may be also formulated through constraints on both components.

In this chapter, we also deal with maximum likelihood estimation of a constrained LM model, which may be performed by an Expectation-Maximization (EM) algorithm having the same structure as that illustrated for the basic LM model. The only relevant differences are in the M-step.

Finally, we deal with model selection and we show how to test a hypothesis formulated by a certain constraint on the parameters through a likelihood ratio statistic between nested models. In certain cases, this statistic has null asymptotic distribution of chi-squared type, whereas in other cases, a more sophisticated distribution arises.

Note that in order to properly describe the different constraints and illustrate them through some examples, we widely use the matrix notation. Then, it is useful to clarify now that by $\mathbf{0}_h$ we denote a column vector of h zeros, by $\mathbf{1}_h$ a column vector of h ones, by $\boldsymbol{O}_{hj}$ an $h \times j$ matrix of zeros, and by $\boldsymbol{I}_h$ the identity matrix of size h. The dimension of these vectors and matrices will not be indicated explicitly when it is clear from the context. Moreover, we use the symbol $\boldsymbol{d}_{hj}$ to denote a column vector of dimension j with all elements equal to zero, apart from the h-th element which is equal to one; for instance, we have $\boldsymbol{d}_{23} = (0, 1, 0)'$. Finally, $\otimes$ denotes the Kronecker product.

4.2 Constraints on the measurement model

In the following, we first consider constraints that may be posed on the measurement model in the case of univariate responses (only one response at each time occasion) and then in the case of multivariate responses (more responses at each time occasion). In the first case the constraints concern the probabilities $\phi_{y|u}^{(t)} = f_{Y^{(t)}|U^{(t)}}(y|u)$, whereas in the second they concern the probabilities $\phi_{jy|u}^{(t)} = f_{Y_j^{(t)}|U^{(t)}}(y|u)$. Note that we are using an extended notation for these conditional response probabilities because we allow that they are time varying, so that we have a *time-heterogeneous* measurement model. However, we suppose that suitable constraints are put on these probabilities so that the resulting model is identifiable. Moreover, in order to compute the manifest probability $f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$, we can use the same recursion as in (3.6), with $\phi_{y^{(t)}|u}$ substituted by $\phi_{y^{(t)}|u}^{(t)}$. The same rule applies for the multivariate case, with each probability $\phi_{\boldsymbol{y}^{(t)}|u}$ substituted by $\phi_{\boldsymbol{y}^{(t)}|u}^{(t)}$ where, in general, $\phi_{\boldsymbol{y}|u}^{(t)} = f_{\boldsymbol{Y}^{(t)}|U^{(t)}}(\boldsymbol{y}|u)$ and, due to the assumption of local independence, we have

$$\phi_{\boldsymbol{y}|u}^{(t)} = \prod_{j=1}^{r} \phi_{jy_j|u}^{(t)}.$$

4.2.1 Univariate formulation

In order to express the constraints of interest, it is convenient to collect the conditional probabilities $\phi_{y|u}^{(t)}$, $y = 0, \dots, c-1$, in the column vector $\boldsymbol{\phi}_u^{(t)}$. The simplest constraint that may be expressed on these probabilities is

$$\boldsymbol{\phi}_u^{(t)} = \boldsymbol{\phi}_u, \quad t = 1, \dots, T,\ u = 1, \dots, k, \tag{4.1}$$

where $\boldsymbol{\phi}_u$ has elements $\phi_{y|u}$, $y = 0, \dots, c-1$. It corresponds to the hypothesis that the measurement model is *time homogeneous*, according to which the distribution of the responses depends only on the corresponding latent variable and there is no dependence on time. This constraint is assumed whenever no other constraints are assumed in order to avoid model unidentifiability. Obviously, if only constraint (4.1) is assumed, the basic LM model results.

More sophisticated constraints may be formulated in the form of a *generalized linear model* (GLM), see McCullagh and Nelder (1989) and

Agresti (2002), as

$$\boldsymbol{\eta}_u^{(t)} = \boldsymbol{W}_u^{(t)}\boldsymbol{\beta}, \tag{4.2}$$

where $\boldsymbol{\eta}_u^{(t)} = \boldsymbol{g}(\boldsymbol{\phi}_u^{(t)})$, with $\boldsymbol{g}(\cdot)$ being a suitable *link function*, and $\boldsymbol{W}_u^{(t)}$ is a design matrix with a number of rows typically equal to $c-1$ and a number of column equal to the dimension of $\boldsymbol{\beta}$. In practice, $\boldsymbol{g}(\cdot)$ is a function that projects the probability vector $\boldsymbol{\phi}_u^{(t)}$ onto the $\mathbb{R}^{c-1}$ space. The choice of this function depends on the nature of the response variables. In the following, we show some typical formulations for binary response variables and for categorical response variables with more than two categories.

4.2.1.1 Binary response variables

With binary response variables ($c = 2$), it is natural to use a *logit link function.* In this case, we have

$$\eta_u^{(t)} = \log \frac{\phi_{1|u}^{(t)}}{\phi_{0|u}^{(t)}}.$$

An interesting model based on this parametrization is the LM Rasch model (Bartolucci et al., 2008), which is illustrated in the following example.

Example 10 — LM Rasch model. *Consider the parametrization*

$$\eta_u^{(t)} = \log \frac{\phi_{1|u}^{(t)}}{\phi_{0|u}^{(t)}} = \alpha_u - \psi^{(t)}, \qquad t = 1,\dots,T, \quad u = 1,\dots,k, \tag{4.3}$$

which may be expressed as in (4.2) by including the parameters α_u and $\psi^{(t)}$ in $\boldsymbol{\beta}$. In particular, taking into account the identifiability constraint $\alpha_1 = 0$, this parametrization may be expressed by letting

$$\boldsymbol{\beta} = (\alpha_2,\dots,\alpha_k,\psi^{(1)},\dots,\psi^{(T)})'$$

and

$$\boldsymbol{W}_u^{(t)} = \begin{cases} (\boldsymbol{0}'_{k-1} \quad -\boldsymbol{d}'_{tT}), & u = 1, \\ (\boldsymbol{d}'_{u-1,k-1} \quad -\boldsymbol{d}'_{tT}), & u = 2,\dots,k. \end{cases}$$

The resulting model may be seen as an extension of the latent class (LC) Rasch model described in Example 1, which makes sense for data derived from the administration of a set of T test items to a group of n subjects. In this case, α_u is interpreted as the ability level of the subjects in latent state u, whereas $\psi^{(t)}$ is interpreted as the difficulty level of *item t. This extension allows the ability level to evolve even during the test administration, possibly due to learning-through-training or tiring phenomena.*

Note that, in the case of item responses administered at the same occasion, we are not properly in a longitudinal context in which the same response variable is repeatedly observed. Nevertheless, the model makes sense as an alternative to the LC Rasch model (De Leeuw and Verhelst, 1986; Lindsay et al., 1991) and to test violation of the latter; see Bartolucci (2006), Bartolucci et al. (2008), and Bartolucci and Solis-Trapala (2010).

An advantage of parametrization (4.3) is that it implies the hypothesis of *monotonicity*, already mentioned in Section 2.5.2, that is,

$$\phi_{1|1}^{(t)} \leq \cdots \leq \phi_{1|k}^{(t)}, \quad t = 1, \dots, T. \tag{4.4}$$

Note that this constraint may be included in the model regardless of a specific parametrization, such as that in (4.3). In particular, the approach of Bartolucci (2006) allows for inequality constraints in the general form

$$\boldsymbol{K}\boldsymbol{\beta} \geq \boldsymbol{0},$$

where $\boldsymbol{\beta}$ is the parameter vector in (4.2). In this way we can formulate a model with ordered latent states in a nonparametric way. In particular, using design matrices $\boldsymbol{W}_u^{(t)}$ formulated so that $\boldsymbol{\beta}$ contains all the logits $\eta_u^{(t)}$, with u running faster than t, the constraint of monotonicity may be expressed by letting

$$\boldsymbol{K} = \boldsymbol{I}_T \otimes \left[\begin{pmatrix} \boldsymbol{0}_{k-1} & \boldsymbol{I}_{k-1} \end{pmatrix} - \begin{pmatrix} \boldsymbol{I}_{k-1} & \boldsymbol{0}_{k-1} \end{pmatrix} \right],$$

where the second matrix at right-hand side (rhs) produces first differences.

Further to the logit link function, in the presence of binary response variables, we can use a *probit link function.* In this case, we have

$$\eta_u^{(t)} = \Phi^{-1}(\phi_{1|u}^{(t)}),$$

where $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative distribution function of the standard normal distribution. This link function is typically motivated by assuming the existence of an underlying continuous variable with an additive Gaussian noise. The response variable is equal to 1 if this variable is greater than a certain threshold, and it is equal to 0 otherwise. A similar representation with logistic error terms leads to the logit link function described above. In general, a probit model is slightly more complex to estimate with respect to the corresponding logit model. For a detailed description of the difference between the two formulations we refer to McCullagh and Nelder (1989).

4.2.1.2 Categorical response variables

In the case of categorical response variables having more than two categories ($c > 2$), a variety of link functions is available. The choice of the specific function depends on the nature of the response variables, essentially ordinal or nonordinal. These are the most important cases:

- With nonordinal response variables, a natural choice is the link function based on *reference category logits*. In this case, the vector $\boldsymbol{\eta}_u^{(t)}$ has elements

$$\eta_{y|u}^{(t)} = \log \frac{\phi_{y|u}^{(t)}}{\phi_{0|u}^{(t)}}, \quad y = 1, \ldots, c-1. \tag{4.5}$$

  Note that the first category of the response variable (0) is taken as the reference category and then it has a special role. However, this choice is irrelevant for the inference, in the sense that the maximum of the likelihood remains unchanged if we take another reference category. On the other hand, an appropriate choice of this category makes the interpretation of the results easier.

- With ordinal response variables, the statistical literature (see, among others, Colombi and Forcina, 2001) suggests the use of logits of type local, global, or continuation. *Local logits*, also known as adjacent-category logits, are formulated so that the elements of $\boldsymbol{\eta}_u^{(t)}$ are

$$\eta_{y|u}^{(t)} = \log \frac{\phi_{y|u}^{(t)}}{\phi_{y-1|u}^{(t)}}, \quad y = 1, \ldots, c-1.$$

  Global logits, which are closely related to cumulative logits, are based on an interesting interpretation based on an underlying continuous outcome which is suitably discretized, but leads to a more complex model. These logits are based on comparing the survival function with the distribution function, leading to

$$\eta_{y|u}^{(t)} = \log \frac{\phi_{y|u}^{(t)} + \cdots + \phi_{c-1|u}^{(t)}}{\phi_{0|u}^{(t)} + \cdots + \phi_{y-1|u}^{(t)}}, \quad y = 1, \ldots, c-1. \tag{4.6}$$

  Note that, by definition, these logits are nonincreasing ordered. Finally, *continuation logits* are of type

$$\eta_{y|u}^{(t)} = \log \frac{\phi_{y|u}^{(t)} + \cdots + \phi_{c-1|u}^{(t)}}{\phi_{y-1|u}^{(t)}}, \quad y = 1, \ldots, c-1.$$

An alternative to the link function based on global logits function is the *ordered probit link function.* In this case, the elements of $\boldsymbol{\eta}_u^{(t)}$ are

$$\eta_{y|u}^{(t)} = \Phi^{-1}(\phi_{y|u}^{(t)} + \cdots + \phi_{c-1|u}^{(t)}), \quad y = 1, \ldots, c-1,$$

and they are nonincreasing ordered. These logits are formulated by assuming the existence of an underlying continuous response variable which involves a Guassian error term, whereas global logits involve a logistic error term. In practice, the global logit function is the counterpart for ordinal variables of the logit link function for binary variables and the ordered probit link function is the counterpart of the probit link function.

In choosing the most suitable link function for a given application, we also have to take into account the simplicity in estimating the resulting model. In general, an interesting feature of all the types of logit illustrated above is that the corresponding link functions may be simply expressed as

$$\boldsymbol{g}(\boldsymbol{\phi}_u^{(t)}) = \boldsymbol{C} \log(\boldsymbol{M}\boldsymbol{\phi}_u^{(t)}), \tag{4.7}$$

where $\boldsymbol{C}$ is a suitable matrix of contrasts, with elements 0, 1 or -1, and $\boldsymbol{M}$ is a marginalization matrix, with elements 0 or 1.

In Appendix 1 we report the definition of the matrices $\boldsymbol{C}$ and $\boldsymbol{M}$ according to the type of logit; see also McCullagh (1980) and Colombi and Forcina (2001). For instance, with $c = 3$ and global logits, we have that

$$\boldsymbol{C} = \begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{M} = \left(\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \hline 0 & 1 & 1 \\ 0 & 0 & 1 \end{array}\right).$$

As we will show, the above formulation of $\boldsymbol{\eta}_u^{(t)}$ is useful in implementing the estimation algorithm. The following example clarifies the formulation of LM models for ordinal variables.

Example 11 — LM Graded Response and Partial Credit models. *Consider a version of the LM Rasch model for ordinal data, which is based on global logits and on the assumption*

$$\eta_{y|u}^{(t)} = \log \frac{\phi_{y|u}^{(t)} + \cdots + \phi_{c-1|u}^{(t)}}{\phi_{0|u}^{(t)} + \cdots + \phi_{y-1|u}^{(t)}} = \alpha_u - \psi_y^{(t)}, \quad y = 1, \ldots, c-1, \tag{4.8}$$

for $t = 1, \ldots, T$ and $u = 1, \ldots, k$. Once the vector $\boldsymbol{\eta}_u^{(t)}$ is constructed as described above, this model is formulated as in (4.2) with

$$\boldsymbol{\beta} = (\alpha_2, \ldots, \alpha_k, \psi_1^{(1)}, \psi_2^{(1)}, \ldots, \psi_{c-1}^{(T)})'$$

and

$$\boldsymbol{W}_u^{(t)} = \begin{cases} \left(\boldsymbol{O}_{k-1,c-1} \quad \boldsymbol{d}'_{tT} \otimes \boldsymbol{I}_{c-1}\right), & u = 1, \\ \left(\boldsymbol{d}'_{u-1,k-1} \otimes \boldsymbol{1}_{c-1} \quad \boldsymbol{d}'_{tT} \otimes \boldsymbol{I}_{c-1}\right), & u = 2,\dots,k. \end{cases}$$

In particular, in the case of item responses, α_u is still interpreted as the ability level of subjects in latent state u, whereas $\psi_y^{(t)}$ is interpreted as the difficulty in achieving at least a level y in responding to item t. The resulting model is an LM version of the Graded Response model (Samejima, 1969).

Note that, by using local instead of global logits, an LM version of the Partial Credit model (Masters, 1982) may be formulated.

Finally, it is worth noting that the constraint of time homogeneity formulated by equation (4.1) may be included into (4.2) by requiring $\boldsymbol{W}_u^{(t)} = \boldsymbol{W}_u$, $t = 1,\dots,T$, $u = 1,\dots,k$. However, when only constraint (4.1) is of interest, it is not advisable to rely on a parametrization based on a link function, since the resulting model would be more complex to estimate.

4.2.2 Multivariate formulation

As clarified in Section 3.3, in the case of multivariate responses an LM model is formulated by assuming the conditional independence of the response variables at the same occasion, that is $Y_1^{(t)},\dots,Y_r^{(t)}$, given the corresponding latent variable $U^{(t)}$. This is an extension of the local independence assumption.

The same principle holds in the case of constraints on the measurement model. This means, in practice, that anyone of the constraints illustrated in the previous section may be adopted for each single response variable. More precisely, let $\boldsymbol{\phi}_{j|u}^{(t)}$ be the vector with elements $\phi_{jy|u}^{(t)}$, $y = 0,\dots,c_j - 1$. Then, the constraint of time homogeneity is formulated, as an extension of constraint (4.1), as follows:

$$\boldsymbol{\phi}_{j|u}^{(t)} = \boldsymbol{\phi}_{j|u}, \quad j = 1,\dots,r,\ t = 1,\dots,T,\ u = 1,\dots,k, \tag{4.9}$$

where $\boldsymbol{\phi}_{j|u}$ is a common vector of conditional response probabilities with elements $\phi_{jy|u}$, $y = 0,\dots,c_j - 1$. This constraint is in general assumed if no other constraints are assumed on the conditional response probabilities.

Moreover, we can adopt a GLM formulation to parameterize the conditional distribution of each response variable as in (4.2). In this case, we have

$$\boldsymbol{\eta}_{j|u}^{(t)} = \boldsymbol{W}_{j|u}^{(t)}\boldsymbol{\beta}, \tag{4.10}$$

where $\boldsymbol{\eta}_{j|u}^{(t)} = \boldsymbol{g}_j(\boldsymbol{\phi}_{j|u}^{(t)})$, with $\boldsymbol{g}_j(\cdot)$ being a link function of the type illustrated in Section 4.2.1. This parametrization may include constraint (4.9) by letting $\boldsymbol{W}_{j|u}^{(t)} = \boldsymbol{W}_{j|u}$.

4.3 Constraints on the latent model

In absence of individual covariates, no interesting constraints may be expressed on the initial probabilities π_u. The only constraint that may be of interest is that of uniform initial probabilities, that is,

$$\pi_u = 1/k, \quad u = 1, \ldots, k. \tag{4.11}$$

This means that, at the beginning of the period of observation, each latent state includes the same proportion of subjects.

More interesting constraints may be expressed on the transition probabilities, which allow us to strongly reduce the number of parameters of the model. In order to express these constraints in a concise form, let $\boldsymbol{\Pi}^{(t)}$ denote the transition probability matrix with elements $\pi_{u|\bar{u}}^{(t)}$, $\bar{u}, u = 1, \ldots, k$, arranged by letting $\bar{u}$ run by row and u run by column.

The simplest constraint on the latent model is

$$\boldsymbol{\Pi}^{(t)} = \boldsymbol{\Pi}, \quad t = 2, \ldots, T, \tag{4.12}$$

where $\boldsymbol{\Pi}$ is a common transition matrix with elements $\pi_{u|\bar{u}}$, $\bar{u}, u = 1, \ldots, k$. It corresponds to the hypothesis that the *Markov chain is time homogenous.*

A weaker version of the above constraint is

$$\boldsymbol{\Pi}^{(t)} = \begin{cases} \boldsymbol{\Pi}^{*(1)}, & t = 2, \ldots, T^*, \\ \boldsymbol{\Pi}^{*(2)}, & t = T^* + 1, \ldots, T, \end{cases} \tag{4.13}$$

with T^* between 2 and $T-1$ and $\boldsymbol{\Pi}^{*(1)}$ and $\boldsymbol{\Pi}^{*(2)}$ being separate transition matrices, with elements $\pi_{u|\bar{u}}^{*(1)}$ and $\pi_{u|\bar{u}}^{*(2)}$, respectively. This corresponds to the hypothesis of *partial time homogeneity* based on two different transition matrices, one until occasion T^* and the other for transitions after this occasion; see Bartolucci et al. (2007) for a model based on this hypothesis.

More sophisticated parametrizations may be formulated by a linear model or a GLM on the transition probabilities. These two different approaches are outlined in the following.

4.3.1 Linear model on the transition probabilities

Linear models have the advantage of permitting the formulation of the constraint that certain probabilities are equal to 0, so that transition between two given states is not possible. In particular, Bartolucci (2006) considered a formulation that in our case may be expressed as

$$\boldsymbol{\rho}_{\bar{u}}^{(t)} = \boldsymbol{Z}_{\bar{u}}^{(t)}\boldsymbol{\delta}, \tag{4.14}$$

where $\boldsymbol{\rho}_{\bar{u}}^{(t)}$ is the column vector containing the elements of the $\bar{u}$-th row of the t-th transition matrix, apart from the diagonal element, that is, $\pi_{u|\bar{u}}^{(t)}$, $u = 1, \ldots, k$, $u \neq \bar{u}$. For instance, with $k = 3$, we have

$$\boldsymbol{\rho}_1^{(t)} = (\pi_{2|1}^{(t)}, \pi_{3|1}^{(t)})', \ \boldsymbol{\rho}_2^{(t)} = (\pi_{1|2}^{(t)}, \pi_{3|2}^{(t)})', \ \boldsymbol{\rho}_3^{(t)} = (\pi_{1|3}^{(t)}, \pi_{2|3}^{(t)})'.$$

Moreover, $\boldsymbol{Z}_{\bar{u}}^{(t)}$ is a design matrix and $\boldsymbol{\delta}$ is a corresponding vector of parameters. In order to ensure that all the transition probabilities are nonnegative, we have to impose suitable restrictions on $\boldsymbol{\delta}$. Due to these restrictions, estimation may be more difficult and we are not in a standard inferential problem; see Section 4.5.2 and Appendix 2 for further details.

The following examples clarify the use of the linear parametrization in (4.14).

Example 12 — Linear constraints on the transition probabilities. *The simplest constraint is that all the off-diagonal elements of the transition matrix* $\boldsymbol{\Pi}^{(t)}$ *are equal to each other; with* $k = 3$, *for instance, we have*

$$\boldsymbol{\Pi}^{(t)} = \begin{pmatrix} 1-2\delta^{(t)} & \delta^{(t)} & \delta^{(t)} \\ \delta^{(t)} & 1-2\delta^{(t)} & \delta^{(t)} \\ \delta^{(t)} & \delta^{(t)} & 1-2\delta^{(t)} \end{pmatrix}, \quad t = 2, \ldots, T. \tag{4.15}$$

This model may be formulated through (4.14), with

$$\boldsymbol{\delta} = (\delta^{(2)}, \ldots, \delta^{(T)})'$$

and

$$\boldsymbol{Z}_{\bar{u}}^{(t)} = \boldsymbol{d}'_{t-1,T-1} \otimes \boldsymbol{1}_{k-1}.$$

In order to also incorporate the constraint of time homogeneity, we set all the design matrices $\boldsymbol{Z}_{\bar{u}}^{(t)}$ *equal to* $\boldsymbol{1}_{k-1}$.

A less stringent constraint is that each transition matrix is symmetric, so that the probability of transition from latent state $\bar{u}$ *to latent state*

u is the same as that of the reverse transition. For $t = 2, \dots, T$, we have

$$\mathbf{\Pi}^{(t)} = \begin{pmatrix} 1-(\delta_1^{(t)}+\delta_2^{(t)}) & \delta_1^{(t)} & \delta_2^{(t)} \\ \delta_1^{(t)} & 1-(\delta_1^{(t)}+\delta_3^{(t)}) & \delta_3^{(t)} \\ \delta_2^{(t)} & \delta_3^{(t)} & 1-(\delta_2^{(t)}+\delta_3^{(t)}) \end{pmatrix}.$$

We can formulate this model by letting

$$\boldsymbol{\delta} = (\delta_1^{(2)}, \delta_2^{(2)}, \dots, \delta_{k(k-1)/2}^{(T)})', \qquad (4.16)$$

with each matrix $\boldsymbol{Z}_{\bar{u}}^{(t)}$ obtained by selecting in a suitable way $k-1$ rows from the matrix $\boldsymbol{d}'_{t-1,T-1} \otimes \boldsymbol{I}_{k(k-1)/2}$.

Finally, when the latent states are ordered in a meaningful way by assuming, for instance, that (4.4) holds, it may be interesting to formulate the hypothesis that a subject in latent state $\bar{u}$ may move only to latent state $u = \bar{u}+1, \dots, k$. With $k = 3$, for instance, we have

$$\mathbf{\Pi}^{(t)} = \begin{pmatrix} 1-(\delta_1^{(t)}+\delta_2^{(t)}) & \delta_1^{(t)} & \delta_2^{(t)} \\ 0 & 1-\delta_3^{(t)} & \delta_3^{(t)} \\ 0 & 0 & 1 \end{pmatrix}, \quad t = 2, \dots, T. \qquad (4.17)$$

In this case the parameter vector $\boldsymbol{\delta}$ is defined as in (4.16), and $\boldsymbol{Z}_{\bar{u}}^{(t)}$ is obtained by selecting $\bar{u}-1$ rows from $\boldsymbol{d}'_{t-1,T-1} \otimes \boldsymbol{I}_{k(k-1)/2}$. A first example of use of the above restriction was provided by Collins and Wugalter (1992), in which latent states correspond to ordered developmental levels. According to the underlying developmental psychology theory, children may make a transition to a next stage but never return to a previous stage.

4.3.2 Generalized linear model on the transition probabilities

An alternative approach to parametrize the transition probabilities is through a suitable link function for each row of the transition matrix. The model may be then formulated as

$$\boldsymbol{\lambda}_{\bar{u}}^{(t)} = \boldsymbol{Z}_{\bar{u}}^{(t)} \boldsymbol{\delta}, \qquad (4.18)$$

with $\boldsymbol{\lambda}_{\bar{u}}^{(t)} = \boldsymbol{h}_{\bar{u}}(\boldsymbol{\pi}_{\bar{u}}^{(t)})$, where $\boldsymbol{\pi}_{\bar{u}}^{(t)}$ is the vector with elements $\pi_{u|\bar{u}}^{(t)}$, $u = 1, \dots, k$, and $\boldsymbol{Z}_{\bar{u}}^{(t)}$ is a suitable design matrix. A possible link function is based on logits with respect to the diagonal element, so that $\boldsymbol{\lambda}_{\bar{u}}^{(t)}$ has $k-1$ elements equal to

$$\lambda_{u|\bar{u}}^{(t)} = \log \frac{\pi_{u|\bar{u}}^{(t)}}{\pi_{\bar{u}|\bar{u}}^{(t)}}, \quad u = 1, \dots, k, \ u \neq \bar{u}. \qquad (4.19)$$

An alternative parametrization, which makes sense with ordered latent states, is based on global logits, so that the elements of each vector $\boldsymbol{\lambda}_{\bar{u}}^{(t)}$ are

$$\lambda_{u|\bar{u}}^{(t)} = \log \frac{\pi_{u|\bar{u}}^{(t)} + \cdots + \pi_{k|\bar{u}}^{(t)}}{\pi_{1|\bar{u}}^{(t)} + \cdots + \pi_{u-1|\bar{u}}^{(t)}}, \quad u = 2, \ldots, k. \tag{4.20}$$

Many other link functions may be formulated, such as that of type ordered probit, in way similar to that illustrated in Section 4.2.1.

It is worth noting that we can combine a parametrization of type (4.18) with the constraint that certain transition probabilities are equal to 0. In this case the link function $\boldsymbol{h}_{\bar{u}}(\cdot)$ must be applied to only those elements in the $\bar{u}$-th row of the t-th transition matrix that are not constrained to be equal to 0. In general, the size of each vector $\boldsymbol{\lambda}_{\bar{u}}^{(t)}$ is equal to the number of these elements minus 1 and the design matrices $\boldsymbol{Z}_{\bar{u}}^{(t)}$ in (4.18) need to be defined accordingly. In any case, we can again express the link function in the form

$$\boldsymbol{h}_{\bar{u}}(\boldsymbol{\pi}_{\bar{u}}^{(t)}) = \boldsymbol{D}_{\bar{u}} \log(\boldsymbol{N}_{\bar{u}} \boldsymbol{\pi}_{\bar{u}}^{(t)}), \tag{4.21}$$

with matrices $\boldsymbol{D}_{\bar{u}}$ and $\boldsymbol{N}_{\bar{u}}$ defined in Appendix 1. The following example clarifies this case.

Example 13 — Tridiagonal transition matrix with logit parametrization. *A strong reduction of the number of parameters may be achieved by assuming that the transition matrices are tridiagonal, so that transition from state $\bar{u}$ is only allowed to state $u = \bar{u} - 1, \bar{u} + 1$; with $k = 4$, for instance, we have*

$$\boldsymbol{\Pi}^{(t)} = \begin{pmatrix} \pi_{1|1}^{(t)} & \pi_{2|1}^{(t)} & 0 & 0 \\ \pi_{1|2}^{(t)} & \pi_{2|2}^{(t)} & \pi_{3|2}^{(t)} & 0 \\ 0 & \pi_{2|3}^{(t)} & \pi_{3|3}^{(t)} & \pi_{4|3}^{(t)} \\ 0 & 0 & \pi_{3|4}^{(t)} & \pi_{4|4}^{(t)} \end{pmatrix}. \tag{4.22}$$

This constraint makes only sense if latent states are suitably ordered. In this case, we may also assume a parametrization based on logits of type (4.19) by requiring, for instance, that

$$\lambda_{u|\bar{u}}^{(t)} = \log \frac{\pi_{u|\bar{u}}^{(t)}}{\pi_{\bar{u}|\bar{u}}^{(t)}} = \delta_{1u} + \delta_{2,(3+u-\bar{u})/2}, \quad t = 2, \ldots, T,\ \bar{u} = 1, \ldots, k,$$

where $u=2$ *for* $\bar{u}=1$, $u=k-1$ *for* $\bar{u}=k$, *and* $u=\bar{u}-1,\bar{u}+1$ *for* $\bar{u}=2,\ldots,k-1$; *in this way, the index* $(3+u-\bar{u})/2$ *may be equal to only 1 or 2. This parametrization may be still expressed as in (4.18) with*

$$\boldsymbol{\delta}=(\delta_{11},\ldots,\delta_{1k},\delta_{21},\delta_{22})'$$

and

$$\boldsymbol{Z}_{\bar{u}}^{(t)}=\begin{cases}\begin{pmatrix}\boldsymbol{d}_{1k}' & \boldsymbol{d}_{22}'\end{pmatrix}, & \bar{u}=1,\\ \begin{pmatrix}\boldsymbol{d}_{\bar{u}k}'\otimes\boldsymbol{1}_2 & \boldsymbol{I}_2\end{pmatrix}, & \bar{u}=2,\ldots,k-1,\\ \begin{pmatrix}\boldsymbol{d}_{kk}' & \boldsymbol{d}_{12}'\end{pmatrix}, & \bar{u}=k.\end{cases}$$

Note that in this case the vector $\boldsymbol{\lambda}_{\bar{u}}^{(t)}$ *contains only one logit for* $\bar{u}=1,k$ *and 2 logits for* $\bar{u}=2,\ldots,k-1$; *see also Appendix 1.*

4.4 Maximum likelihood estimation

Under the constraints illustrated in Sections 4.2 and 4.3, maximum likelihood estimation of the parameters of the LM model is carried out by the same EM algorithm illustrated in Section 3.5, in which we only have to modify the M-step according to the constraint of interest. In the following we describe this algorithm in detail.

We recall that the observed data consist of the vectors $\tilde{\boldsymbol{y}}_i$, $i=1,\ldots,n$, which are independent realizations of the random vector $\tilde{\boldsymbol{Y}}$. Moreover, the model log-likelihood may still be expressed as

$$\ell(\boldsymbol{\theta})=\sum_{i=1}^{n}\log f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}}_i), \tag{4.23}$$

where $\boldsymbol{\theta}$ is the vector of all model parameters arranged in a suitable way. The structure of this parameter vector is different according to the model formulation. In particular:

- if only constraint (4.1) (or (4.9) in the multivariate case) is posed on the measurement model, the parameter vector includes the conditional response probabilities $\phi_{y|u}$ (or $\phi_{jy|u}$ in the multivariate case) which are common to all time occasions. Otherwise, if a GLM is assumed on these probabilities, $\boldsymbol{\theta}$ contains the vector $\boldsymbol{\beta}$ in (4.2) (or in (4.10) in the multivariate case);

- if no constraints are posed on the latent model, $\boldsymbol{\theta}$ also contains the initial probabilities π_u and the transition probabilities $\pi_{u|\bar{u}}^{(t)}$. If instead constraint (4.11) is assumed, then the initial probabilities are not included, whereas if constraint (4.12) or (4.13) is assumed, the common transition probabilities $\pi_{u|\bar{u}}$ or $\pi^{*(1)}_{u|\bar{u}}$ and $\pi^{*(2)}_{u|\bar{u}}$ are included instead of the probabilities $\pi^{(t)}_{u|\bar{u}}$. Finally, the latter probabilities are substituted by the vector $\boldsymbol{\delta}$ if a linear model of type (4.14) or a GLM of type (4.18) is assumed.

We recall that an equivalent expression for the log-likelihood in (4.23) is

$$\ell(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{y}}} \log f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}}),$$

where $n_{\tilde{\boldsymbol{y}}}$ is the frequency of the response configuration $\tilde{\boldsymbol{y}}$. This is the expression that we use in practice since it leads to a more efficient computation of the maximum likelihood estimate.

4.4.1 Expectation-Maximization algorithm

As for the basic LM model, the complete data are represented by the pairs $(\boldsymbol{u}_i, \tilde{\boldsymbol{y}}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{u}_i$ stands for the sequence of latent states of subject i. These data are considered in the complete data log-likelihood on which the EM algorithm is based. Details are provided in the following subsections for the case of univariate and multivariate responses.

4.4.1.1 Univariate formulation

In the univariate case, the complete data log-likelihood has the following expression:

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) &= \sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{y=0}^{c-1} a^{(t)}_{uy} \log \phi^{(t)}_{y|u} \\ &+ \sum_{u=1}^{k} b^{(1)}_{u} \log \pi_u + \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k} b^{(t)}_{\bar{u}u} \log \pi^{(t)}_{u|\bar{u}}, \qquad (4.24)\end{aligned}$$

where the frequencies $a^{(t)}_{uy}$, $b^{(t)}_{u}$, and $b^{(t)}_{\bar{u}u}$ are defined as in Section 3.5.1.1.

Based on the above log-likelihood, the EM algorithm is performed by alternating two steps until convergence:

- **E-step**: compute the expected value of the frequencies in (4.24), given the observed data and the current value of the parameters. In the univariate case, these expected values are computed through (3.13), (3.14), and (3.15) and are denoted by $\hat{a}^{(t)}_{uy}$, $\hat{b}^{(t)}_{u}$, and $\hat{b}^{(t)}_{\bar{u}u}$. Note that these rules are based on recursion (3.19), in which, however, each probability $\phi_{y^{(t+1)}|u}$ is substituted by $\phi^{(t+1)}_{y^{(t+1)}|u}$.

- **M-step**: update the parameter vector by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$. As usual, we can separately update the blocks of parameters as follows:

  - *Conditional response probabilities*: if the constraint of time homogeneity of these probabilities is assumed, see equation (4.1), then they are updated through (3.16) as for the basic LM model. If instead a GLM of type (4.2) is assumed on these probabilities, then an iterative algorithm is necessary to update the parameter vector $\boldsymbol{\beta}$ on which the response probabilities depend. This algorithm is briefly illustrated in Appendix 2.
  - *Initial probabilities*: if these probabilities are unconstrained, then they are updated as in (3.17). Under the constraint of uniform initial probabilities, the M-step simply skips the update of these parameters, whose values are fixed as in (4.11).
  - *Transition probabilities*: as above, if the transition probabilities are unconstrained, then they are updated in the usual way; see equation (3.18). When the Markov chain is assumed to be time homogenous, as in (4.12), these probabilities are updated as follows:

$$\pi_{u|\bar{u}} = \frac{\sum_{t=2}^{T} \hat{b}_{\bar{u}u}^{(t)}}{\sum_{t=2}^{T} \hat{b}_{\bar{u}}^{(t-1)}}, \quad \bar{u}, u = 1, \dots, k. \tag{4.25}$$

A similar rule may be applied in the partial homogeneity case to update the parameters $\pi_{u|\bar{u}}^{*(1)}$ and $\pi_{u|\bar{u}}^{*(2)}$ in (4.13). For the parameters of the first type, the sum $\sum_{t=2}^{T}$ in the equation above needs to be substituted by $\sum_{t=2}^{T^*}$; for those of second type, this sum must be substituted by $\sum_{t=T^*+1}^{T}$. Finally, an iterative algorithm is necessary for the case of a linear model or a GLM on the transition probabilities; see Appendix 2 for details. Note however that, for certain hypotheses formulated by a linear model, an explicit way to update the parameters is available.

Regarding the convergence of the EM algorithm, the same considerations at the end of Section 3.5.1.1 may be drawn even for the case of constrained LM models on which this chapter is focused. In particular, in order to decide when to stop the algorithm, rules (3.22) and (3.23) may still be applied.

4.4.1.2 Multivariate formulation

In the multivariate case, the complete data log-likelihood has the following expression:

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) &= \sum_{j=1}^{r}\sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{y=0}^{c-1} a_{juy}^{(t)} \log \phi_{jy|u}^{(t)} \\ &+ \sum_{u=1}^{k} b_u^{(1)} \log \pi_u + \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k} b_{\bar{u}u}^{(t)} \log \pi_{u|\bar{u}}^{(t)},\end{aligned}$$

where $a_{juy}^{(t)}$ is defined as in (3.24).

The EM algorithm is based on the same steps outlined above for the univariate case. The main difference is that, at the E-step, the expected frequencies $\hat{a}_{juy}^{(t)}$ are computed instead of $\hat{a}_{uy}^{(t)}$; see Section 3.5.1.2.

Regarding the M-step, if only the constraint of time homogeneity of the conditional response probabilities is assumed, see (4.9), then they are updated through (3.25) as for the basic LM model. If instead a GLM of type (4.10) is assumed on these probabilities, then an iterative algorithm is necessary to update the parameter vector $\boldsymbol{\beta}$ on which the response probabilities depend. This algorithm is outlined in Appendix 2.

4.4.1.3 Initialization of the algorithm and model identifiability

For the initialization of the EM algorithm, we suggest to proceed along the same lines as in Section 3.5.1.3. In particular, we strongly advise using a combination of a deterministic method and of a stochastic method and then to choose the best solution obtained by these methods as the maximum likelihood estimate, denoted by $\hat{\boldsymbol{\theta}}$.

The deterministic rule to initialize the EM algorithm is based on formulating starting values for the conditional response probabilities $\phi_{y|u}^{(t)}$ (or $\phi_{jy|u}^{(t)}$ in the multivariate case), the initial probabilities π_u, and the transition probabilities $\pi_{u|\bar{u}}^{(t)}$. Then, the EM algorithm proceeds as described above by updating the parameter vector $\boldsymbol{\theta}$. In the univariate case, in particular, we first compute the empirical global logits

$$\eta_y^{(t)} = \log \frac{\sum_{i=1}^{n} I(y_i^{(t)} \geq y)}{\sum_{i=1}^{n} I(y_i^{(t)} < y)}, \quad t = 1,\ldots,T,\ y = 1,\ldots,c-1.$$

Then, each conditional probability $\phi_{y|u}^{(t)}$ is initialized by the same transformation of these logits introduced in Section 3.5.1.3 and applied for

$t = 1, \ldots, t$ and $y = 1, \ldots, c-1$; see equation (3.26). A similar rule is applied in the presence of multivariate responses. Moreover, the initial and transition probabilities are initialized as in (3.27) and (3.28). Similarly, the random starting rule we suggest is essentially the same illustrated in Section 3.5.1.3, which is based on random numbers generated from a uniform distribution between 0 and 1. The only difference is that we have to assign a distinct random value to the conditional response probabilities for $t = 1, \ldots, T$.

It is worth noting that the initialization method illustrated above, in both its deterministic and stochastic versions, has the advantage of not being related to a specific parametrization because it is based on choosing starting values for the probabilities $\phi^{(t)}_{y|u}$ (or $\phi^{(t)}_{jy|u}$), π_u, and $\pi^{(t)}_{u|\bar{u}}$ of the unconstrained LM model with the same number of latent states. Moreover, through the same initialization methods and along the same lines as in Section 3.5.1.3 we check for model identifiability even for constrained LM models.

4.5 Model selection and hypothesis testing

In applying a constrained LM model, a fundamental issue is that of model selection. In this case, we have to select both the number of latent states, if this number is not *a priori* defined, and the specific constraints on the measurement and/or the latent model. Obviously, this problem is strongly related to that of testing hypotheses which are expressed through constraints of the same type. These issues are discussed in more detail in the following.

4.5.1 Model selection

The strategy that we suggest for model selection is based on first selecting the number of states k, if this choice is necessary, and then the specific constraints of interest. In particular, the preferred model selection for k is the Bayesian information criterion (BIC) defined in Section 3.6, even if we acknowledge that alternative criteria are available, the application of which depends on the specific context of application. We apply this criterion to the basic LM model illustrated in the previous chapter, which does not include constraints of interest.

Once the number of states is selected for the basic LM model, we suggest taking this as the reference model and including constraints of interest in the measurement or in the latent model. Obviously, these constraints must correspond to hypotheses which make sense in the context of application. For instance, constraint (4.8), which gives rise to a longitudinal version of the Graded Response model described in Example 11, makes sense with ordinal response variables only. The model to be selected is the one with the smallest BIC index, provided that this is the adopted model selection criterion.

In principle, in order to select the best model we need to try all the possible combinations of constraints of interest on the measurement and the latent model. However, especially when the number of possible constraints is large and the model fitting is computationally intensive, it may be more convenient to sequentially introduce these constraints and retain the constraint that, at each attempt, leads to a reduction of the BIC index. Alternatively, we can retain this constraint if the corresponding hypothesis is not rejected by a likelihood ratio test when this test can be validly applied, as described in the following section. In this regard, one must be aware that introducing the constraints in a different sequence might lead to different final models. Therefore, a sensible strategy must be chosen in advance. For instance, we may initially introduce the constraints which are of most interest, or those leading to the strongest reduction in the number of parameters, and then the other possible constraints.

4.5.2 Hypothesis testing

As usual, a hypothesis H_0 of interest, which is formulated by a constraint of the type illustrated in Sections 4.2 and 4.3, may be tested by the likelihood ratio statistic

$$LR = -2(\hat{\ell}_0 - \hat{\ell}_1),$$

where $\hat{\ell}_1$ is the maximum value of the likelihood of the unconstrained model and $\hat{\ell}_0$ is that of the restricted model. The same test may be applied within the sequential model selection procedure described above, although testing hypotheses is often related to a scientific interest which not necessarily corresponds to a model selection problem. For a discussion on this point see Burnham and Anderson (2002, Section 6.9).

When the usual regularity conditions hold, the null asymptotic distribution of the above test statistic is of chi-squared type with a number of degrees of freedom equal to the number of nonredundant constraints used to formulate H_0. The latter is typically equal to the difference in the number of free parameters between the two models that are compared. These regularity conditions hold for most of the constraints formulated in this section. For instance, in the case of binary response variables, through LR we can test the hypothesis that the conditional probabilities of success follow a Rasch model; see equation (4.3). In this case, the statistic LR has a chi-squared null asymptotic distribution with $kT - [T + (k-1)]$ degrees of freedom, where $T + (k-1)$ is the number of free parameters involved in (4.3).

The main case where the usual regularity conditions do not hold is when H_0 is formulated by a linear model on the transition probabilities, such as that certain of these probabilities are constrained to be equal to zero; see equation (4.14). In this case, a boundary problem occurs (Self and Liang, 1987) and, as proved by Bartolucci (2006), the null asymptotic distribution is of chi-bar-squared type. This distribution corresponds to a mixture of chi-squared distributions with weights which may be computed by explicit formulae or estimated by a simple Monte Carlo method; see Shapiro (1988) and Silvapulle and Sen (2004, Chapter 3). The weights depend on the information matrix for the model parameters.

An interesting result is when we assume that the transition matrices depend on only one parameter, as in (4.15) with $\delta^{(t)} = \delta$, $t = 2, \ldots, T$. In this case, the hypothesis $H_0 : \delta = 0$ may be tested by the likelihood ratio LR, whose null asymptotic distribution corresponds to a mixture between 0 and a chi-squared distribution with one degree of freedom. The weights of this mixture are 0.5 and 0.5. This means that the p-value for testing this hypothesis may be explicitly computed as

$$p\text{−value} = \frac{1}{2} p(\chi_1^2 \geq LR),$$

that is, as one half the p-value that we would compute if we were in a regular inferential problem.

4.6 Applications

In this section, we show how constrained LM models may be used for the analysis of the two datasets already analyzed by the basic LM model in the previous chapter.

4.6.1 Marijuana consumption dataset

For the Marijuana consumption dataset, we first consider an LM model which is based on only the following assumptions:

- the conditional response probabilities satisfy the parametrization

$$\eta_{y|u}^{(t)} = \alpha_u + \psi_y, \quad t = 1, \ldots, T,\ u = 1, \ldots, k,\ y = 1, \ldots, c-1, \quad (4.26)$$

  where $\eta_{y|u}^{(t)}$ is the y-th global logit for the t-th response variable, given that the subject is in the u-th latent state; see definition (4.6);

- the transition probabilities are time homogeneous as in (4.12).

This model is denoted by M_6. Then, we consider the further constraint that the transition matrix is tridiagonal as in (4.22). The model resulting by including this hypothesis in model M_6 is here denoted by M_7. Finally, we also consider the hypothesis that the transition matrix is upper triangular as in (4.17). When included into model M_6, model M_8 results.

Note that formulation (4.26) is based on one parameter for each latent state and another for each response category, apart from the first. The parameters of the first type, in particular, measure the tendency to use marijuana and then allow us to order the latent states according to this tendency. Moreover, the parameters associated with the response categories are time constant because the response variables correspond to repeated measurements of the same phenomenon under the same circumstances. Therefore, the evolution of the marijuana consumption is represented only by the latent process and, in particular, by the transition probabilities. Under this parametrization the hypothesis that the transition matrix is tridiagonal is completely reasonable, as the hypothesis that this matrix is upper triangular.

Using the same number of latent states selected in Section 3.7.1, the maximum log-likelihoods of models M_6, M_7, and M_8 are shown in Table 4.1 in comparison to that of the basic LM model, denoted by M_5. The table also shows the corresponding number of parameters and the Akaike information criterion (AIC) and BIC indices.

TABLE 4.1
Maximum log-likelihood, number of parameters, and AIC and BIC indices for models M_5 to M_8

Model	$\hat{\ell}$	#par	*AIC*	*BIC*
M_5	−646.89	32	1357.79	1468.76
M_6	−659.59	12	1343.18	1384.81
M_7	−660.60	10	1341.20	1375.89
M_8	−661.93	9	1341.85	1373.07

On the basis of these results, we conclude that model M_6 has to be preferred to the basic LM model (M_5) because the first has much lower values of AIC and BIC indices. Moreover, the deviance between the two models is $LR = 25.40$ with 20 degrees-of-freedom, so that a p-value of 0.186 results. In order to compute this p-value we relied on a standard asymptotic method, because the usual regularity conditions hold.

Once model M_6 is adopted, with a strong reduction of the number of parameters with respect to the basic LM model, we can test the hypothesis that the transition matrix is tridiagonal and the hypothesis that this matrix is upper triangular. Testing these hypothesis amounts to compute the deviance, with respect to model M_6, of model M_7 and model M_8, respectively. In the first case we have $LR = 2.02$, and in the second we have $LR = 4.67$. In order to compute a p-value for these deviances we have to rely on a chi-bar-squared distribution, as illustrated in Section 4.5.2. For the first deviance a p-value of 0.172 results, whereas for the second deviance a p-value of 0.059 results. Therefore, both hypotheses cannot be rejected, but the first seems to be more supported by the data. However, note that a direct comparison between the two hypotheses is not possible by a likelihood ratio statistic because neither of them is nested in the other. On the other hand, the deviance for the hypothesis that the transition matrix is diagonal is $LR = 233.73$ with a p-value smaller than 0.001.

On the basis of the above results, Bartolucci (2006) suggested model M_7 as a sensible model for the marijuana consumption dataset. Note that this model has the smallest value of the AIC index, but not the smallest BIC index, confirming that the two criteria may lead to choosing different models, especially with large samples.

The parameter estimates obtained under model M_7 are reported in Tables 4.2 and 4.3. The first table, in particular, reports the estimates of the parameters in (4.26) involved in the measurement model, whereas the second table reports the estimates of the initial probabilities and of the transition probabilities.

The estimates in Table 4.2 show that the latent states are ordered according to the tendency of marijuana consumption. These estimates

TABLE 4.2
Estimates of the parameters α_u and ψ_y under model M_7

u	$\hat{\alpha}_u$	y	$\hat{\psi}_y$
1	0.000	1	0.165
2	5.751	2	0.686
3	10.876		

TABLE 4.3
Estimates of the initial probabilities π_u and transition probabilities $\pi_{u|\bar{u}}$ under model M_7

			$\hat{\pi}_{u\|\bar{u}}$		
u	$\hat{\pi}_u$	$\bar{u}$	$u=1$	$u=2$	$u=3$
1	0.896	1	0.835	0.165	0.000
2	0.089	2	0.070	0.686	0.244
3	0.015	3	0.000	0.082	0.918

may be easily converted into conditional response probabilities by the rules outlined in Appendix 1. Such probabilities are displayed in Table 4.4.

According to the estimates in Table 4.3, most subjects are in the first state at the beginning of the period of observation with only very few in the third state. Moreover, there is a general tendency toward an increase of the marijuana consumption because each upper diagonal element is higher than the corresponding symmetric element; that is, the probability of transition from state 1 to state 2 is higher than the probability of the reverse transition (0.165 versus 0.070) and the same happens for the transition from state 2 to state 3 (0.244 versus 0.082).

Obviously, even for a constrained model we can obtain the marginal distribution of the latent states for every time occasion through recursion (3.8). The results are provided in Table 4.5 and confirm the increasing tendency of marijuana consumption mentioned above and already noted in Section 3.7.1 on the basis of the basic LM model (M_5). In fact, these probabilities are very close to those in Table 3.6, confirming in this way that, despite its parsimony, model M_7 leads to conclusions very similar to those based on model M_5.

TABLE 4.4
Estimates of the conditional response probabilities $\phi_{y|u}^{(t)}$ under model M_7

	$\hat{\phi}_{y\|u}^{(t)}$		
u	$y=0$	$y=1$	$y=2$
1	0.9935	0.3229	0.0028
2	0.0063	0.5960	0.0602
3	0.0003	0.0811	0.9370

TABLE 4.5
Estimated marginal distribution of the latent states for each time occasion under model M_7

	$\hat{f}_{U^{(t)}}(u)$		
t	$u = 1$	$u = 2$	$u = 3$
1	0.8961	0.0893	0.0146
2	0.7548	0.2100	0.0352
3	0.6452	0.2711	0.0837
4	0.5580	0.2989	0.1431
5	0.4872	0.3085	0.2044

4.6.2 Criminal conviction history dataset

For the analysis of this dataset, the structure of which is described in Section 1.4.2, we consider an LM model based on the constraint that both the conditional response probabilities and the transition probabilities are time homogeneous, so that both (4.9) and (4.12) hold. This model is denoted by M_7. We also consider models in which the transition probabilities are partially time homogenous, as in (4.13), for different values of T^* between 2 and 5, whereas the conditional response probabilities still satisfy constraint (4.9). This type of model is denoted by M_8.

The results in terms of maximum log-likelihood, and AIC and BIC indices from the fitting of the above models are reported in Table 4.6 in comparison to the basic LM model, which is denoted by M_6 as in Section 3.7.2. All models are fitted with $k = 6$ latent states.

These results show that the model including the hypothesis of time homogeneity of the transition probabilities (M_7) is not preferable to the

TABLE 4.6
Maximum log-likelihood, number of parameters, and AIC and BIC indices for models M_6, M_7, and M_8 with different values of T^* for the partial time-homogeneity constraint on the transition probabilities

model	T^*	$\hat{\ell}$	#par	*AIC*	*BIC*
M_6	–	−110676.09	215	221782.19	223689.57
M_7	–	−111331.45	95	222852.90	223695.70
M_8	2	−110761.77	125	221773.54	222882.49
M_8	3	−110930.24	125	222110.47	223219.42
M_8	4	−110986.77	125	222223.54	223332.49
M_8	2	−110952.28	125	222154.57	223263.51

TABLE 4.7
Estimates of the conditional response probabilities $\phi_{j1|u}$ under model M_8 with $T^* = 2$

	$\hat{\phi}_{j1\|u}$					
j	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$	$u = 6$
1	0.0015	0.0000	0.2317	0.0461	0.0450	0.3798
2	0.0006	0.0016	0.0246	0.0130	0.0114	0.0369
3	0.0000	0.0122	0.0223	0.0673	0.4533	0.5629
4	0.0000	0.0000	0.0087	0.0044	0.0146	0.0671
5	0.0035	0.0731	0.1400	0.6563	0.6768	0.8293
6	0.0008	0.0000	0.0172	0.3266	0.0121	0.2403
7	0.0008	0.0077	0.1359	0.0278	0.1321	0.3274
8	0.0007	0.0000	0.1090	0.0346	0.0000	0.1892
9	0.0000	0.0000	0.0069	0.0226	0.0000	0.0870
10	0.0004	0.0000	0.1068	0.0999	0.1708	0.4816

basic LM model (M_6), since it has larger values of both AIC and BIC indices. On the other hand, for the partial time-homogeneity model (M_8) with $T^* = 2$, we have the lowest values of AIC and BIC indices; this is the model we select. Note that the hypothesis underlying this model is that the subjects move between latent states in a different way when they are in the first age band with respect to when they are in other age bands.

Under the selected model M_8, we obtain the maximum likelihood parameter estimates shown in Tables 4.7, 4.8, and 4.9, which are structured like the corresponding tables in Section 3.7.2.

On the basis of the estimates in Table 4.7, we reach substantially the same conclusions as in Section 3.7.2 about the interpretation of the latent states in terms of typology of crime. In fact, the estimates in this table are very similar to those in Table 3.8.

On the other hand, the estimates of the initial probabilities in Table 4.8 are rather different with respect to those obtained under the basic

TABLE 4.8
Estimates of the initial probabilities π_u under model M_8 with $T^* = 2$

u	$\hat{\pi}_u$
1	0.6908
2	0.2769
3	0.0005
4	0.0003
5	0.0307
6	0.0007

TABLE 4.9
Estimates of the transition probabilities $\pi^{*(t)}_{u|\bar{u}}$ under model M_8 with $T^* = 2$

		$\hat{\pi}^{*(t)}_{u\|\bar{u}}$					
T^*	$\bar{u}$	$u=1$	$u=2$	$u=3$	$u=4$	$u=5$	$u=6$
1	1	0.8384	0.1528	0.0088	0.0000	0.0000	0.0000
	2	0.1005	0.7059	0.0811	0.0438	0.0514	0.0174
	3	0.1950	0.0000	0.5150	0.2901	0.0000	0.0000
	4	0.4775	0.0000	0.5225	0.0000	0.0000	0.0000
	5	0.0000	0.2643	0.0989	0.0000	0.3096	0.3272
	6	0.0877	0.0000	0.0000	0.0000	0.0000	0.9123
2	1	0.9986	0.0000	0.0000	0.0014	0.0000	0.0000
	2	0.4819	0.4126	0.0646	0.0385	0.0002	0.0022
	3	0.4900	0.1070	0.3764	0.0185	0.0000	0.0081
	4	0.4022	0.3022	0.0393	0.2538	0.0000	0.0025
	5	0.0000	0.4796	0.1660	0.0615	0.2033	0.0897
	6	0.0000	0.1175	0.2693	0.0783	0.0781	0.4568

LM model (see Table 3.9). In particular, under model M_8 we estimate a smaller presence of subjects in the first latent state, that of nonoffenders, at the beginning of the period of observation.

Finally, on the basis of the estimated transition probabilities in Table 4.9, we can see the difference between the evolution of the criminal behavior from the first to the second age band and from the second to the other age bands. In particular, the first transition matrix corresponds to a higher level of persistence than the second transition matrix. In fact, for the first four rows of the latter matrix the highest element is in the first column, meaning that subjects tend to considerably reduce the tendency to commit crimes as they become older. This feature was already noted by Bartolucci et al. (2007) in analyzing the same dataset by an LM model which also takes into account the covariate gender; this application will be illustrated in detail in the next chapter.

Appendix 1: Marginal parametrization

Parametrization of the conditional response probabilities

According to the chosen parametrization, the matrices $\boldsymbol{C}$ and $\boldsymbol{M}$ in (4.7) are defined as follows. In the case of reference-category logits, as in (4.5), $\boldsymbol{C} = (-\boldsymbol{1}_{c-1} \quad \boldsymbol{I}_{c-1})$ and $\boldsymbol{M} = \boldsymbol{I}_c$. In the other cases, we have $\boldsymbol{C} = (-\boldsymbol{I}_{c-1} \quad \boldsymbol{I}_{c-1})$ and

$$\boldsymbol{M} = \begin{cases} \begin{pmatrix} \boldsymbol{I}_{c-1} & \boldsymbol{0}_{c-1} \\ \boldsymbol{0}_{c-1} & \boldsymbol{I}_{c-1} \end{pmatrix}, & \text{for logits of type local,} \\ \begin{pmatrix} \boldsymbol{T}_{c-1} & \boldsymbol{0}_{c-1} \\ \boldsymbol{0}_{c-1} & \boldsymbol{T}'_{c-1} \end{pmatrix}, & \text{for logits of type global,} \\ \begin{pmatrix} \boldsymbol{I}_{c-1} & \boldsymbol{0}_{c-1} \\ \boldsymbol{0}_{c-1} & \boldsymbol{T}'_{c-1} \end{pmatrix}, & \text{for logits of type continuation,} \end{cases}$$

where $\boldsymbol{T}_h$ is an $h \times h$ lower triangular matrix of ones.

An important issue is how to transform a certain vector of marginal parameters $\boldsymbol{\eta}_u^{(t)}$ back to the vector probabilities $\boldsymbol{\phi}_u^{(t)}$. For reference-category logits, which are based on the simplest transformation in which $\boldsymbol{M} = \boldsymbol{I}_c$, we have that

$$\boldsymbol{\phi}_u^{(t)} = \frac{1}{\boldsymbol{1}'_c \exp(\boldsymbol{G}\boldsymbol{\eta}_u^{(t)})} \exp(\boldsymbol{G}\boldsymbol{\eta}_u^{(t)}), \quad \boldsymbol{G} = \begin{pmatrix} \boldsymbol{0}'_{c-1} \\ \boldsymbol{I}_{c-1} \end{pmatrix}. \tag{4.27}$$

The same transformation holds for local logits, but with

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{0}'_{c-1} \\ \boldsymbol{T}_{c-1} \end{pmatrix}.$$

For the other types of logit, it is first convenient to obtain the survival function and then the single probabilities $\phi_{y|u}^{(t)}$ in a way similar to that adopted to initialize the EM algorithm; see Section 3.5.1.3. In particular, let

$$\phi_{y|u}^{*(t)} = p(Y^{(t)} \geq y | U^{(t)} = u), \quad y = 0, \ldots, c.$$

For the global logit link function we have

$$\phi_{y|u}^{*(t)} = \begin{cases} 1, & y = 0, \\ \frac{\exp(\eta_{y|u}^{(t)})}{1+\exp(\eta_{y|u}^{(t)})}, & y = 1, \ldots, c-1, \\ 0, & y = c, \end{cases} \tag{4.28}$$

whereas for the continuation logit link function we have

$$\phi_{y|u}^{*(t)} = \begin{cases} 1, & y = 0, \\ \prod_{h=1}^{y} \frac{\exp(\eta_{h|u}^{(t)})}{1+\exp(\eta_{h|u}^{(t)})}, & y = 1, \ldots, c-1, \\ 0, & y = c. \end{cases}$$

Once the above cumulate probabilities are available, the elements of $\boldsymbol{\phi}_u^{(t)}$ are computed by the differences

$$\phi_{y|u}^{(t)} = \phi_{y+1|u}^{*(t)} - \phi_{y|u}^{*(t)}, \quad y = 0, \ldots, c-1.$$

Parametrization of the transition probabilities

For the parametrization in (4.21), there are essentially two cases to consider. In the first case, leading to the logits in (4.19), the matrix $\boldsymbol{D}_{\bar{u}}$, $\bar{u} = 1, \ldots, k$, is obtained by removing the $\bar{u}$-th row from the matrix $\boldsymbol{I}_k - \boldsymbol{d}'_{\bar{u}k} \otimes \boldsymbol{1}_k$, whereas $\boldsymbol{N}_{\bar{u}} = \boldsymbol{I}_k$. The second case, leading to the global logits in (4.20), is formulated with $\boldsymbol{D}_{\bar{u}}$ and $\boldsymbol{N}_{\bar{u}}$ defined as above, that is,

$$\boldsymbol{D}_{\bar{u}} = \begin{pmatrix} -\boldsymbol{I}_{k-1} & \boldsymbol{I}_{k-1} \end{pmatrix} \quad \text{and} \quad \boldsymbol{N}_{\bar{u}} = \begin{pmatrix} \boldsymbol{T}_{k-1} & \boldsymbol{0}_{k-1} \\ \boldsymbol{0}_{k-1} & \boldsymbol{T}'_{k-1} \end{pmatrix}, \quad u = 1, \ldots, k.$$

When parametrization (4.19) or (4.20) is applied to only certain elements in each vector $\boldsymbol{\pi}_{\bar{u}}^{(t)}$, while the other elements are constrained to 0, then $\boldsymbol{N}_{\bar{u}}$ is obtained by selecting a subset of rows from the matrix defined as above. The matrix $\boldsymbol{D}_{\bar{u}}$ is defined accordingly, so that it has a number of rows equal to the number of elements of $\boldsymbol{\pi}_{\bar{u}}^{(t)}$, which are not constrained to 0, minus 1 and a number of columns equal to the number of rows of $\boldsymbol{N}_{\bar{u}}$. For instance, for the case of the triangular transition matrix (4.22) and the first type of parametrization, we have

$$\boldsymbol{D}_{\bar{u}} = \begin{cases} (-1, 1), & \bar{u} = 1, \\ \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}, & \bar{u} = 2, \ldots, k-1, \\ (-1, 1), & \bar{u} = k, \end{cases}$$

and

$$\boldsymbol{N}_{\bar{u}} = \begin{cases} (\boldsymbol{I}_2 \quad \boldsymbol{O}_{2,k-2}), & \bar{u} = 1, \\ \left(\boldsymbol{O}_{3,\bar{u}-2} \quad \boldsymbol{I}_3 \quad \boldsymbol{O}_{3,k-\bar{u}-1}\right), & \bar{u} = 2, \ldots, k-1, \\ (\boldsymbol{O}_{2,k-2} \quad \boldsymbol{I}_2), & \bar{u} = k. \end{cases}$$

To transform $\boldsymbol{\lambda}_{\bar{u}}^{(t)}$ back into the probability vector $\boldsymbol{\pi}_{\bar{u}}^{(t)}$, we have to apply rules similar to those illustrated above for the conditional response probabilities. Starting from the case that no elements of $\boldsymbol{\pi}_{\bar{u}}^{(t)}$ are constrained to 0, when logits of type (4.19) are assumed, then $\boldsymbol{\pi}_{\bar{u}}^{(t)}$ may be computed by the same transformation in (4.27) with $\boldsymbol{G}$ obtained by removing the $\bar{u}$-th column from the matrix $\boldsymbol{I}_k$. Moreover, if logits of type (4.20) are used, then the same method of inversion based on (4.28) can be adopted. Finally, when certain elements of $\boldsymbol{\pi}_{\bar{u}}^{(t)}$ are constrained to 0, these transformations are applied to obtain only the remaining elements of this probability vector.

Appendix 2: Implementation of the M-step

Parametrization of the conditional response probabilities

Under a parametrization of type (4.2), the parameter vector $\boldsymbol{\beta}$ is updated by maximizing the corresponding component of the expected value of the complete data log-likelihood that, with univariate responses, is given by

$$\hat{\ell}_1^*(\boldsymbol{\beta}) = \sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{y=0}^{c-1} \hat{a}_{uy}^{(t)} \log \phi_{y|u}^{(t)} = \sum_{t=1}^{T}\sum_{u=1}^{k} (\hat{\boldsymbol{a}}_u^{(t)})' \log \boldsymbol{\phi}_u^{(t)},$$

where $\hat{\boldsymbol{a}}_u^{(t)}$ is a column vector with elements $\hat{a}_{uy}^{(t)}$, $y = 0, \ldots, c-1$, and $\boldsymbol{\phi}_u^{(t)}$ is defined accordingly on the basis of the probabilities $\phi_{y|u}^{(t)}$.

The maximization of $\hat{\ell}_1^*(\boldsymbol{\beta})$ may be performed by a standard iterative algorithm, such as the Fisher-scoring, which is based on the score vector and the expected information matrix for this log-likelihood function. These quantities have specific expressions depending on the adopted link function which, in turn, depends on the nature of the response variables. However, the formulation of the link function in (4.7) leads to the following general expressions:

$$\hat{\boldsymbol{s}}_1^*(\boldsymbol{\beta}) = \sum_{t=1}^{T}\sum_{u=1}^{k} (\boldsymbol{W}_u^{(t)})'(\boldsymbol{R}_u^{(t)})'\boldsymbol{G}'(\hat{\boldsymbol{a}}_u^{(t)} - \hat{b}_u^{(t)}\boldsymbol{\phi}_u^{(t)}), \qquad (4.29)$$

$$\hat{\boldsymbol{F}}_1^*(\boldsymbol{\beta}) = \sum_{t=1}^{T}\sum_{u=1}^{k} \hat{b}_u^{(t)}(\boldsymbol{W}_u^{(t)})'(\boldsymbol{R}_u^{(t)})'\boldsymbol{G}'\boldsymbol{\Omega}_u^{(t)}\boldsymbol{G}\boldsymbol{R}_u^{(t)}\boldsymbol{W}_u^{(t)}, \qquad (4.30)$$

where $\boldsymbol{W}_u^{(t)}$ is the design matrix in (4.2), $\boldsymbol{G}$ is defined in (4.27), and $\boldsymbol{R}_u^{(t)}$ is a suitable derivative matrix obtained as

$$\boldsymbol{R}_u^{(t)} = [\boldsymbol{C}(\boldsymbol{M}\boldsymbol{\phi}_u^{(t)})^{-1}\boldsymbol{M}\mathrm{diag}(\boldsymbol{\phi}_u^{(t)})\boldsymbol{G}]^{-1}$$

depending on the matrices $\boldsymbol{C}$ and $\boldsymbol{M}$ used in the adopted link function; see Bartolucci et al. (2001) and Colombi and Forcina (2001) for details. Moreover, we have $\boldsymbol{\Omega}_u^{(t)} = [\mathrm{diag}(\boldsymbol{\phi}_u^{(t)}) - \boldsymbol{\phi}_u^{(t)}(\boldsymbol{\phi}_u^{(t)})']$. In practice, the Fisher-scoring algorithm updates the parameter vector $\boldsymbol{\beta}$ by adding at each step $[\hat{\boldsymbol{F}}_1^*(\boldsymbol{\beta})]^{-1}\hat{\boldsymbol{s}}_1^*(\boldsymbol{\beta})$ until convergence. See also Bartolucci (2006) for a more detailed description of the above algorithm which takes into account constraints of type $\boldsymbol{K\beta} \geq \boldsymbol{0}$.

The same iterative algorithm as above may be used in the multivariate case, when a parametrization of type (4.10) is adopted and then the following function needs to be maximized:

$$\hat{\ell}_1^*(\boldsymbol{\beta}) = \sum_{j=1}^{r}\sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{y=0}^{c-1} \hat{a}_{juy}^{(t)} \log \phi_{jy|u}^{(t)}.$$

Parametrization of the transition probabilities

In order to update the parameter vector $\boldsymbol{\delta}$ in (4.14) or (4.18), we have to maximize the corresponding component of the expected value of the complete data log-likelihood, that is,

$$\hat{\ell}_3^*(\boldsymbol{\delta}) = \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k} \hat{b}_{\bar{u}u}^{(t)} \log \pi_{u|\bar{u}}^{(t)} = \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k} (\hat{\boldsymbol{b}}_{\bar{u}}^{(t)})' \log \boldsymbol{\pi}_{\bar{u}}^{(t)},$$

where $\hat{\boldsymbol{b}}_{\bar{u}}^{(t)}$ is a column vector with elements $\hat{b}_{\bar{u}u}^{(t)}$, $u = 1,\dots,k$, and $\boldsymbol{\pi}_{\bar{u}}^{(t)}$ is defined accordingly on the basis of the probabilities $\pi_{u|\bar{u}}^{(t)}$. In computing $\hat{\ell}_3^*(\boldsymbol{\delta})$, we have to pay attention to the presence of transition probabilities equal to 0; this is one of the constraints that may be of interest. Therefore, we adopt the convention that $\hat{b}_{\bar{u}u}^{(t)} \log \pi_{u|\bar{u}}^{(t)} \equiv 0$ whenever $\pi_{u|\bar{u}}^{(t)}$ is constrained to be equal to 0 or $\hat{b}_{\bar{u}u}^{(t)} = 0$.

When a GLM of type (4.18) is assumed on the transition probabilities, the same Fisher-scoring algorithm as above may be used to maximize $\hat{\ell}_3^*(\boldsymbol{\delta})$. This algorithm is based on the score vector and the Fisher information matrix with respect to $\boldsymbol{\delta}$ having expressions similar to (4.29) and (4.30). The only relevant adjustment that is necessary is for the presence of possible transition probabilities equal to 0.

For the case of a linear model on the transition probabilities, formulated through (4.14), we can use a similar algorithm. In this case, however, we have to take into account that the parameter vector $\boldsymbol{\delta}$ is subjected to specific inequality constraints. These constraints on $\boldsymbol{\delta}$ are necessary to ensure that the corresponding transition probabilities satisfy the usual constraints, that is,

$$\pi_{u|\bar{u}}^{(t)} \geq 0, \quad t = 2,\dots,T,\ \bar{u}, u = 1,\dots,k,$$

and

$$\sum_{u=1}^{k} \pi_{u|\bar{u}}^{(t)} = 1, \quad t = 2,\dots,T,\ \bar{u} = 1,\dots,k.$$

We refer to Bartolucci (2006) for details about the implementation of this algorithm.

It is worth noting that, for many linear models of interest, an iterative algorithm is not necessary to update the parameter vector $\boldsymbol{\delta}$ at the M-step because an explicit solution is available. This happens, for instance, when each transition matrix $\mathbf{\Pi}^{(t)}$ depends on only one parameter as in (4.15). For this parameter we have an explicit solution given by

$$\delta^{(t)} = \frac{\sum_{\bar{u}=1}^{k} \sum_{u=1, u\neq\bar{u}}^{k} \hat{b}_{\bar{u}u}^{(t)}}{n}, \quad t = 2, \dots, T.$$

Similarly, for the case of upper transition matrices, see (4.17), we have the following explicit solutions for $t = 2, \dots, T$:

$$\delta_1^{(t)} = \frac{\hat{b}_{12}^{(t)}}{\hat{b}_1^{(t-1)}}, \quad \delta_2^{(t)} = \frac{\hat{b}_{13}^{(t)}}{\hat{b}_1^{(t-1)}}, \quad \delta_3^{(t)} = \frac{\hat{b}_{23}^{(t)}}{\hat{b}_{22}^{(t)} + \hat{b}_{23}^{(t)}}.$$

This solution holds for $k = 3$, but it can be easily generalized to large values of k.

5

Including individual covariates and relaxing basic model assumptions

5.1 Introduction

In this chapter, we deal with the inclusion of individual covariates into the basic latent Markov (LM) model. These covariates may be included in the measurement model or in the latent model. This is possible by adopting essentially the same parametrizations used to formulate constraints on the basic LM model; see, in particular, Sections 4.2 and 4.3.

In addition to the inclusion of individual covariates, this chapter deals with relaxing the assumption of local independence and the assumption that the latent Markov chain is of first order. To relax the assumption of local independence, we have to allow the response variables to be dependent even conditionally on the latent process. There are two possibilities in this regard. The first makes sense with both univariate and multivariate responses and consists of allowing each response variable to depend on the lagged response variables (*serial dependence*). This extension is strongly related to that of the inclusion of individual covariates in the measurement model, because both extensions rely on suitable parametrizations of the conditional distribution of the response variables given the latent process. This extension allows us to formulate a model which includes the so-called *state dependence effect* (Heckman, 1981), that is, the effect that experiencing a certain situation at a certain occasion has on the probability of experiencing the same situation in the future, once all the other observable and unobservable explanatory factors have been accounted for.

The second possibility to relax local independence makes sense only with multivariate responses and consists of allowing the response variables at the same time occasion to be dependent even conditionally on the corresponding latent variable (*contemporary dependence*).

Concerning the assumption that the latent Markov chain is of first order, we show alternative models based on higher order Markov chains, discussing in particular the second-order LM model. We will show how this model may be reformulated as a first-order model with suitable constraints on the initial and transition probabilities, so that estimation may be carried out in a rather simple way.

5.2 Notation

Before illustrating in detail how to include individual covariates in an LM model, it is useful to clarify the notation that will be adopted and how it is related to the notation in the previous chapters.

First of all, we introduce the symbol $\boldsymbol{X}^{(t)}$ to denote the vector of individual covariates which are available at the t-th time occasion, $t = 1, \ldots, T$. Moreover, we let $\tilde{\boldsymbol{X}}$ denote the vector of all the individual covariates which is obtained by stacking the vectors $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$. However, we use the same notation as in the previous chapters for the response variables. Then, in the univariate case we have a response variable $Y^{(t)}$ for each time occasion $t = 1, \ldots, T$, whereas in the multivariate case, we have a vector of response variables $\boldsymbol{Y}^{(t)}$ for each of these occasions; this vector has elements $Y_j^{(t)}$, $j = 1, \ldots, r$. These response variables are collected in the vector $\tilde{\boldsymbol{Y}}$, which has T elements in the univariate case and rT in the multivariate case. Finally, the latent variables are still denoted by $U^{(t)}$ and are collected in the vector $\boldsymbol{U}$.

When an LM model is extended to include individual covariates, the assumption of local independence and the assumption that the latent Markov chain is of first order are maintained. However, these assumptions are formulated conditionally on all the covariates collected in $\tilde{\boldsymbol{X}}$.

In order to illustrate the possible model formulations in the presence of individual covariates, we denote the conditional response probabilities by

$$\phi_{y|u\boldsymbol{x}}^{(t)} = f_{Y^{(t)}|U^{(t)},\boldsymbol{X}^{(t)}}(y|u,\boldsymbol{x}), \quad y = 0, \ldots, c-1,$$

in the univariate case and by

$$\phi_{jy|u\boldsymbol{x}}^{(t)} = f_{Y_j^{(t)}|U^{(t)},\boldsymbol{X}^{(t)}}(y|u,\boldsymbol{x}), \quad j = 1, \ldots, r,\ y = 0, \ldots, c_j - 1,$$

in the multivariate case, where $t = 1, \ldots, T$ and $u = 1, \ldots, k$. Accordingly, we have the following initial and transition probabilities of the latent process:

$$\begin{aligned} \pi_{u|\boldsymbol{x}} &= f_{U^{(t)}|\boldsymbol{X}^{(t)}}(u|\boldsymbol{x}), \quad u = 1, \ldots, k \\ \pi_{u|\bar{u}\boldsymbol{x}}^{(t)} &= f_{U^{(t)}|U^{(t-1)},\boldsymbol{X}^{(t)}}(u|\bar{u},\boldsymbol{x}), \quad t = 2, \ldots, T,\ \bar{u}, u = 1, \ldots, k. \end{aligned}$$

In the above expressions, $\boldsymbol{x}$ denotes a realization of $\boldsymbol{X}^{(t)}$, y denotes a realization of $Y^{(t)}$ or $Y_j^{(t)}$, u denotes a realization of $U^{(t)}$, and $\bar{u}$ denotes a realization of $U^{(t-1)}$.

At this stage, it is important to note that the manifest distribution of the response variables corresponds to the conditional distribution of $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{X}}$. Its probability mass function is denoted by $f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}})$, where $\tilde{\boldsymbol{x}}$ denotes a realization of the vector of all the covariates $\tilde{\boldsymbol{X}}$. This probability mass function is related to that of the conditional distribution of $\boldsymbol{U}$ given $\tilde{\boldsymbol{X}}$, denoted by $f_{\boldsymbol{U}|\tilde{\boldsymbol{X}}}(\boldsymbol{u}|\tilde{\boldsymbol{x}})$, and to that of the conditional distribution of $\tilde{\boldsymbol{Y}}$ given $\boldsymbol{U}$ and $\tilde{\boldsymbol{X}}$, denoted by $f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U},\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\boldsymbol{u},\tilde{\boldsymbol{x}})$. These probability functions are defined as follows (the last two expressions hold for the univariate case):

$$\begin{aligned}
f_{\boldsymbol{U}|\tilde{\boldsymbol{X}}}(\boldsymbol{u}|\tilde{\boldsymbol{x}}) &= \pi_{u^{(1)}|\boldsymbol{x}^{(1)}} \prod_{t=2}^{T} \pi^{(t)}_{u^{(t)}|u^{(t-1)}\boldsymbol{x}^{(t)}}, && (5.1)\\
f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U},\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\boldsymbol{u},\tilde{\boldsymbol{x}}) &= \prod_{t=1}^{T} \phi^{(t)}_{y^{(t)}|u^{(t)}\boldsymbol{x}^{(t)}}, && (5.2)\\
f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) &= \sum_{\boldsymbol{u}} \pi_{u^{(1)}|\boldsymbol{x}^{(1)}} \pi^{(2)}_{u^{(2)}|u^{(1)}\boldsymbol{x}^{(2)}} \cdots \pi^{(T)}_{u^{(T)}|u^{(T-1)}\boldsymbol{x}^{(T)}} \\
&\quad \times \phi^{(1)}_{y^{(1)}|u^{(1)}\boldsymbol{x}^{(1)}} \cdots \phi^{(T)}_{y^{(T)}|u^{(T)}\boldsymbol{x}^{(T)}}, && (5.3)
\end{aligned}$$

which are extended versions of (3.2), (3.3), and (3.4), respectively. Accordingly, for the multivariate case we have an expression which extends (3.9). In both cases, the manifest probability of $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{X}}$ may be computed by a forward recursion based on (3.6) and (3.7), with π_u substituted by $\pi_{u|\boldsymbol{x}^{(1)}}$, $\pi^{(t)}_{u|\bar{u}}$ substituted by $\pi^{(t)}_{u|\bar{u}\boldsymbol{x}^{(t)}}$, and $\phi^{(t)}_{y^{(t)}|u}$ substituted by $\phi^{(t)}_{y^{(t)}|u\boldsymbol{x}^{(t)}}$ in the univariate case or $\phi^{(t)}_{\boldsymbol{y}^{(t)}|u\boldsymbol{x}^{(t)}}$ in the multivariate case, where $\phi^{(t)}_{\boldsymbol{y}|u\boldsymbol{x}} = f_{\boldsymbol{Y}^{(t)}|U^{(t)},\boldsymbol{X}^{(t)}}(\boldsymbol{y}|u,\boldsymbol{x})$. The assumption of local independence implies that the last probability may be computed as

$$\phi^{(t)}_{\boldsymbol{y}|u\boldsymbol{x}} = \prod_{j=1}^{r} \phi^{(t)}_{jy_j|u\boldsymbol{x}}. \qquad (5.4)$$

With the necessary adjustments, we can also implement this recursion in matrix notation, as described in Appendix 1 of Chapter 3.

As mentioned above, the assumption of local independence may be relaxed in two different ways; the first is based on allowing serial dependence between the responses. In particular, we consider only the case in which the distribution of each occasion-specific response variable is affected by the response variable (or variables in the multivariate case) observed at the previous time occasion. In this case, we assume that the initial observation $Y^{(0)}$, or the vector of initial observations $\boldsymbol{Y}^{(0)}$ in the multivariate case, is available. Then, this case may be cast into that of individual covariates by including $Y^{(t-1)}$ (or $\boldsymbol{Y}^{(t-1)}$) into the vector $\boldsymbol{X}^{(t)}$ for $t = 1, \ldots, T$. Consequently, by manifest distribution of the response variables we mean their conditional distribution given $Y^{(0)}$ (or $\boldsymbol{Y}^{(0)}$) and the available covariates; the notation needs then to be suitably extended.

The notation needs to be extended even to deal with the case of contemporary dependence between the response variables and to deal with LM models based on a latent higher-order Markov chain. However, we prefer to provide a detailed description of this notation when we illustrate these extensions.

5.3 Covariates in the measurement model

In the following, we show how to include individual covariates in the measurement model. As usual, in order to simplify the exposition, we distinguish between the case of univariate responses and that of multivariate responses.

5.3.1 Univariate formulation

Given that the response variables are categorical, a natural way to include individual covariates in the measurement model is through a generalized linear model parametrization which recalls that used in (4.2) to formulate constraints on this model. This is shown in detail in the following.

Let $\boldsymbol{\phi}_{u\boldsymbol{x}}^{(t)}$ be the vector with elements $\phi_{y|u\boldsymbol{x}}^{(t)}$, $y = 0, \ldots, c-1$, and $\boldsymbol{\eta}_{u\boldsymbol{x}}^{(t)} = \boldsymbol{g}(\boldsymbol{\phi}_{u\boldsymbol{x}}^{(t)})$ a transformation of this vector based on the link function $\boldsymbol{g}(\cdot)$. We assume, for all the configurations $\boldsymbol{x}$ of $\boldsymbol{X}^{(t)}$, the following parametrization:

$$\boldsymbol{\eta}_{u\boldsymbol{x}}^{(t)} = \boldsymbol{W}_{u\boldsymbol{x}}^{(t)}\boldsymbol{\beta}, \quad t = 1, \ldots, T,\ u = 1, \ldots, k, \tag{5.5}$$

where $\boldsymbol{W}_{u\boldsymbol{x}}^{(t)}$ is a design matrix which depends on $\boldsymbol{x}$.

In practice, the same link functions considered in Section 4.2 may be adopted. The choice essentially depends on the nature of the response variables. We recall that, with binary variables, we can adopt either the logit or the probit link function. With response variables having more than two categories, we have a wider choice. In particular, we typically choose a link function based on multinomial logits with nonordinal response variables. With ordinal response variables, we can choose a link function based on local, global, or continuation logits, or we can choose an ordered probit link function. Note that generalized logit link functions may be expressed as in (4.7) on the basis of the matrices $\boldsymbol{C}$ and $\boldsymbol{M}$, the construction of which is clarified in Appendix 1 of Chapter 4.

A parametrization of type (5.5), assumed to include individual covariates in the measurement model, is typically combined with the constraint that the same latent models hold for all subjects. This constraint may be expressed as

$$\pi_{u|\boldsymbol{x}} = \pi_u, \quad u = 1, \ldots, k, \tag{5.6}$$

and

$$\pi_{u|\bar{u}\boldsymbol{x}}^{(t)} = \pi_{u|\bar{u}}^{(t)}, \quad t = 2, \ldots, T,\ \bar{u}, u = 1, \ldots, k, \tag{5.7}$$

for all possible covariate configurations $\boldsymbol{x}$, where π_u and $\pi_{u|\bar{u}}$ are initial and transition probabilities common to all units. This constraint may be formulated together with some of the constraints illustrated in Section 4.3.

Given the similarity with the material in Section 4.2, we do not add further details and conclude this section with two examples.

Example 14 — Logit model with time-varying random effects. *Consider the LM model for binary response variables based on the assumption*

$$\log \frac{\phi_{1|u\boldsymbol{x}}^{(t)}}{\phi_{0|u\boldsymbol{x}}^{(t)}} = \alpha_u + \boldsymbol{x}'\boldsymbol{\psi}, \quad t = 1, \ldots, T,\ u = 1, \ldots, k, \tag{5.8}$$

where α_u is an intercept corresponding to latent state u. Assume also that constraints (5.6) and (5.7) hold. The resulting model may be seen as a logit model with random effects having a discrete distribution and, in particular, being time varying since their value may dynamically evolve during the period of observation. In this way we generalize the model formulated in Example 2.

It is rather easy to show that assumption (5.8) may be formulated as in (5.5) with

$$\boldsymbol{W}_{u\boldsymbol{x}}^{(t)} = \begin{pmatrix} \boldsymbol{d}_{uk}' & \boldsymbol{x}' \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} (\alpha_1, \ldots, \alpha_k) & \boldsymbol{\psi}' \end{pmatrix}', \tag{5.9}$$

where $\boldsymbol{d}_{uk}$ is defined as in Section 4.1.

Example 15 — Proportional odds model with time-varying random effects. *In the case of ordinal response variables with c categories, from* 0 *to* $c-1$*, consider the following parametrization based on global logits*

$$\log \frac{\phi^{(t)}_{y|u\boldsymbol{x}} + \cdots + \phi^{(t)}_{c-1|u\boldsymbol{x}}}{\phi^{(t)}_{0|u\boldsymbol{x}} + \cdots + \phi^{(t)}_{y-1|u\boldsymbol{x}}} = \alpha_{uy} + \boldsymbol{x}'\boldsymbol{\psi}, \tag{5.10}$$

for $t=1,\ldots,T$*,* $u=1,\ldots,k$*, and* $y=1,\ldots,c-1$*, with* α_{uy} *being an intercept which is specific of the category and of the latent state. Assume also that constraints (5.6) and (5.7) hold, together with constraint (4.12). The resulting model may be seen as a random-effects version of the proportional odds model (McCullagh, 1980), in which these random effects are time varying, but differently from the previous example, the corresponding Markov chain is time homogeneous.*

Finally, parametrization (5.10) is a particular case of (5.5) with

$$\boldsymbol{W}^{(t)}_{u\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{d}'_{uk} \otimes \boldsymbol{I}_{c-1} & \boldsymbol{1}_{c-1} \otimes \boldsymbol{x}' \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{k,c-1}) & \boldsymbol{\psi}' \end{pmatrix}',$$

where $\boldsymbol{1}_{c-1}$ *and* $\boldsymbol{I}_{c-1}$ *are defined as in Section 4.1.*

5.3.2 Multivariate formulation

If we retain the assumption of local independence, the inclusion of individual covariates in the measurement model is formulated by a parametrization of type (5.5) for each single response variable. This is in agreement with what has already been discussed in Section 4.2.2 about the formulation of constraints on the measurement model in the multivariate case.

In order to clarify the point above, let $\boldsymbol{\phi}^{(t)}_{j|u\boldsymbol{x}}$ be the vector with elements $\phi^{(t)}_{jy|u\boldsymbol{x}}$, $y=0,\ldots,c_j-1$; also let $\boldsymbol{\eta}^{(t)}_{j|u\boldsymbol{x}} = \boldsymbol{g}_j(\boldsymbol{\phi}^{(t)}_{j|u\boldsymbol{x}})$ be a transformation of this vector based on the link function $\boldsymbol{g}_j(\cdot)$. Then, we assume that

$$\boldsymbol{\eta}^{(t)}_{j|u\boldsymbol{x}} = \boldsymbol{W}^{(t)}_{j|u\boldsymbol{x}}\boldsymbol{\beta},$$

for $j=1,\ldots,r$, $t=1,\ldots,T$, and $u=1,\ldots,k$, where $\boldsymbol{W}^{(t)}_{j|u\boldsymbol{x}}$ is a design matrix depending on the covariates in $\boldsymbol{x}$.

5.4 Covariates in the latent model

A natural way to allow the initial and transition probabilities of the latent Markov chain to depend on the individual covariates is by adopting a parametrization which recalls that in (4.18).

For the initial probabilities, in particular, we assume

$$\boldsymbol{\lambda}_{\boldsymbol{x}} = \boldsymbol{Z}_{\boldsymbol{x}}\boldsymbol{\gamma}, \tag{5.11}$$

where $\boldsymbol{\lambda}_{\boldsymbol{x}} = \boldsymbol{h}(\boldsymbol{\pi}_{\boldsymbol{x}})$, with $\boldsymbol{\pi}_{\boldsymbol{x}} = (\pi_{1|\boldsymbol{x}}, \ldots, \pi_{k|\boldsymbol{x}})'$, $\boldsymbol{Z}_{\boldsymbol{x}}$ is a design matrix depending on the covariates in $\boldsymbol{x}$, and $\boldsymbol{\gamma}$ is the corresponding parameter vector. Similarly, for the transition probabilities we have

$$\boldsymbol{\lambda}_{\bar{u}\boldsymbol{x}}^{(t)} = \boldsymbol{Z}_{\bar{u}\boldsymbol{x}}^{(t)}\boldsymbol{\delta}, \tag{5.12}$$

with $\boldsymbol{\lambda}_{\bar{u}\boldsymbol{x}}^{(t)} = \boldsymbol{h}_{\bar{u}}(\boldsymbol{\pi}_{\bar{u}\boldsymbol{x}})$, where $\boldsymbol{\pi}_{\bar{u}\boldsymbol{x}} = (\pi_{1|\bar{u}\boldsymbol{x}}, \ldots, \pi_{k|\bar{u}\boldsymbol{x}})'$, $\boldsymbol{Z}_{\bar{u}\boldsymbol{x}}^{(t)}$ is a suitable design matrix, and $\boldsymbol{\delta}$ is the corresponding vector of parameters. In the above expressions, $\boldsymbol{h}(\cdot)$ and $\boldsymbol{h}_{\bar{u}}(\cdot)$, $\bar{u} = 1 \ldots, k$, are link functions that may be formulated on the basis on different types of logit. Typically, we use multinomial logits or global logits. The reference category of the multinomial logits is the first category when modeling the initial probabilities and category $\bar{u}$ when modeling the transition probabilities. In the latter case, an expression similar to (4.19) results. Global logits, based on an expression similar to (4.20) are used when the latent states are ordered on the basis of a suitable parametrization of the conditional distribution of the response variables given the latent process. Note that we can combine a parametrization of type (5.12) with the constraint that certain transition probabilities are equal to 0, as in (4.22).

The following example, which is related to Example 13, helps to clarify the possible ways to include the individual covariates in the latent model.

Example 16 — Tridiagonal transition matrix with logit parametrization. *Assuming that the latent states are ordered, we may require that the initial probabilities depend on the covariates in $\boldsymbol{X}^{(1)}$ through a global logits link function as follows:*

$$\lambda_{u|\boldsymbol{x}} = \log \frac{\pi_{u|\boldsymbol{x}} + \cdots + \pi_{k|\boldsymbol{x}}}{\pi_{1|\boldsymbol{x}} + \cdots + \pi_{u-1|\boldsymbol{x}}} = \gamma_{1u} + \boldsymbol{x}'\boldsymbol{\gamma}_2, \quad u = 2, \ldots, k.$$

Moreover, it is reasonable to assume that transition matrices are tridiagonal with transition probabilities parametrized as follows:

$$\lambda_{u|\bar{u}\boldsymbol{x}}^{(t)} = \log \frac{\pi_{u|\bar{u}\boldsymbol{x}}^{(t)}}{\pi_{\bar{u}|\bar{u}\boldsymbol{x}}^{(t)}} = \delta_{1\bar{u}} + \delta_{2,(3+u-\bar{u})/2} + \boldsymbol{x}'\boldsymbol{\delta}_3, \quad t = 2, \ldots, T, \bar{u} = 1, \ldots, k,$$

where $u=2$ *for* $\bar{u}=1$*,* $u=k-1$ *for* $\bar{u}=k$*, and* $u=\bar{u}-1,\bar{u}+1$ *for* $\bar{u}=2,\ldots,k-1$*.*

This parametrization of the initial probabilities may be expressed as in (5.11) with

$$\boldsymbol{\gamma}=\begin{pmatrix}(\gamma_{12},\ldots,\gamma_{1k}) & \boldsymbol{\gamma}_2'\end{pmatrix}' \quad \textit{and} \quad \boldsymbol{Z}_{\boldsymbol{x}}=\begin{pmatrix}\boldsymbol{I}_{k-1} & \boldsymbol{1}_{k-1}\otimes\boldsymbol{x}'\end{pmatrix}.$$

Moreover, the model on the transition probabilities may be formulated as in (5.12), with

$$\boldsymbol{\delta}=((\delta_{11},\ldots,\delta_{1k},\delta_{21},\delta_{22}) \quad \boldsymbol{\delta}_3')'$$

and

$$\boldsymbol{Z}_{\bar{u}\boldsymbol{x}}^{(t)}=\begin{cases}\begin{pmatrix}\boldsymbol{d}_{1k}' & \boldsymbol{d}_{22}' & \boldsymbol{x}'\end{pmatrix}, & \bar{u}=1,\\ \begin{pmatrix}\boldsymbol{d}_{\bar{u}k}'\otimes\boldsymbol{1}_2 & \boldsymbol{I}_2 & \boldsymbol{1}_2\otimes\boldsymbol{x}'\end{pmatrix}, & \bar{u}=2,\ldots,k-1,\\ \begin{pmatrix}\boldsymbol{d}_{kk}' & \boldsymbol{d}_{12}' & \boldsymbol{x}'\end{pmatrix}, & \bar{u}=k.\end{cases}$$

We have to clarify that, when covariates are assumed to enter in the latent model, they are excluded from the measurement model and then, in the univariate case, we have

$$\phi_{y|u\boldsymbol{x}}^{(t)}=\phi_{y|u}^{(t)}, \quad t=1,\ldots,T,\ u=1,\ldots,k,\ y=0,\ldots,c-1,$$

whereas in the multivariate case, we have

$$\phi_{jy|u\boldsymbol{x}}^{(t)}=\phi_{jy|u}^{(t)}, \quad j=1,\ldots,r,t=1,\ldots,T,u=1,\ldots,k,y=0,\ldots,c_j-1,$$

for all covariate configurations $\boldsymbol{x}$. In the above expressions, $\phi_{y|u}^{(t)}$ and $\phi_{jy|u}^{(t)}$ are common to all sample units and may be constrained in a suitable way, as shown in Section 4.2.

5.5 Interpretation of the resulting models

We introduced two different schemes for including individual covariates in an LM model. We suggest to adopt only one scheme, that is to include the covariates in the measurement model (under the constraints $\pi_{u|\boldsymbol{x}}=\pi_u$, $\pi_{u|\bar{u}\boldsymbol{x}}^{(t)}=\pi_{u|\bar{u}}^{(t)}$ for all $\boldsymbol{x}$) or, alternatively, in the latent model (under the constraint $\phi_{y|u\boldsymbol{x}}^{(t)}=\phi_{y|u}^{(t)}$ or $\phi_{jy|u\boldsymbol{x}}^{(t)}=\phi_{jy|u}^{(t)}$ for all $\boldsymbol{x}$). We advise against allowing the covariates to simultaneously affect both the distribution of the latent process and the conditional distribution of the response variables given this process. In fact, the two extensions have distinct interpretations. Moreover, the resulting LM model would be in general of difficult interpretation and its estimation would often be cumbersome.

As already mentioned in Section 1.1, we have to clarify that, when covariates are included in the measurement model, the latent variables are seen as a way to account for the unobserved heterogeneity, that is, the heterogeneity between subjects that we cannot explain on the basis of the observable covariates. The advantage with respect to a standard random-effects model or a latent class model with covariates is that we admit that the effect of unobservable covariates has its own dynamics; for a more complete discussion see Bartolucci and Farcomeni (2009).

When the covariates are included in the latent model, we typically suppose that observable outcomes indirectly measure a latent trait, such as the health condition of elderly people, which may evolve over time. In such a case, the main interest is in modeling the effect of covariates on the latent trait distribution, as in Bartolucci et al. (2009).

5.6 Maximum likelihood estimation

In the presence of individual covariates, the observed data correspond to the vectors $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{y}}_i$, for $i = 1, \ldots, n$. In particular, $\tilde{\boldsymbol{x}}_i$ may be decomposed into the time-specific subvectors of covariates $\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(T)}$. Similarly, we recall that $\tilde{\boldsymbol{y}}_i$ is made up of the elements $y_i^{(1)}, \ldots, y_i^{(T)}$, in the univariate cases, or the subvectors $\boldsymbol{y}_i^{(1)}, \ldots, \boldsymbol{y}_i^{(T)}$, in the multivariate case.

Using the above notation, we have the following expression for the model log-likelihood:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{x}}_i),$$

where $\boldsymbol{\theta}$ is the vector of the parameters which has a structure similar to the one described in Section 4.4. The equivalent expression that is computationally more convenient is

$$\ell(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{x}}} \sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} \log f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}),$$

where $n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}}$ is the joint frequency of the covariate configuration $\tilde{\boldsymbol{x}}$ and the response configuration $\tilde{\boldsymbol{y}}$. The double sum $\sum_{\tilde{\boldsymbol{x}}} \sum_{\tilde{\boldsymbol{y}}}$ is over all the pairs $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ of covariate and response configurations observed at least once.

The likelihood function can be maximized by an Expectation-Maximization (EM) algorithm having a structure very similar to that outlined in Section 4.4.1.

5.6.1 Expectation-Maximization algorithm

The EM algorithm is based on the complete data log-likelihood that, in the univariate case, has expression

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) &= \sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}}\sum_{y=0}^{c-1} a_{u\boldsymbol{x}y}^{(t)} \log \phi_{y|u\boldsymbol{x}}^{(t)} \\ &+ \sum_{u=1}^{k}\sum_{\boldsymbol{x}} b_{u\boldsymbol{x}}^{(1)} \log \pi_{u|\boldsymbol{x}} + \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k}\sum_{\boldsymbol{x}} b_{\bar{u}u\boldsymbol{x}}^{(t)} \log \pi_{u|\bar{u}\boldsymbol{x}}^{(t)}, \qquad (5.13)\end{aligned}$$

where $b_{u\boldsymbol{x}}^{(t)}$ is a frequency of the latent state u and covariate configuration $\boldsymbol{x}$ at occasion t; with reference to the same occasion and covariate configuration, $b_{\bar{u}u\boldsymbol{x}}^{(t)}$ is the number of transitions from state $\bar{u}$ to state u, whereas $a_{u\boldsymbol{x}y}^{(t)}$ is the number of subjects that are in latent state u and provide response y. In the multivariate case, we have a similar expression which depends on the probabilities $\phi_{jy|u\boldsymbol{x}}^{(t)}$.

Based on the complete data log-likelihood, the two steps of the EM algorithm are implemented as follows:

- **E-step**: compute the conditional expected value of each frequency in (5.13); these expected values are denoted by $\hat{a}_{u\boldsymbol{x}y}^{(t)}$, $\hat{b}_{u\boldsymbol{x}}^{(t)}$, and $\hat{b}_{\bar{u}u\boldsymbol{x}}^{(t)}$ and may be computed by modified versions of (3.13), (3.14), and (3.15). In particular, we have

$$\begin{aligned}\hat{a}_{u\boldsymbol{x}y}^{(t)} &= \sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} f_{U^{(t)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) I(\boldsymbol{x}^{(t)}=\boldsymbol{x}, y^{(t)}=y), \\ \hat{b}_{u\boldsymbol{x}}^{(t)} &= \sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} f_{U^{(t)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) I(\boldsymbol{x}^{(t)}=\boldsymbol{x}), \\ \hat{b}_{\bar{u}u\boldsymbol{x}}^{(t)} &= \sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} f_{U^{(t-1)},U^{(t)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(\bar{u},u|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) I(\boldsymbol{x}^{(t)}=\boldsymbol{x}).\end{aligned}$$

  The above posterior probabilities may be computed by applying the same recursions illustrated in Section 3.5; see in particular (3.20) and (3.21).

- **M-step**: maximize the complete data log-likelihood expressed as in (5.13), with each frequency substituted by the corresponding expected value. How to maximize this function depends on the specific formulation of the model and, in particular, on whether the covariates are included in the measurement model or in the latent model. We illustrate this point below, closely following the scheme adopted in Section 4.4.1.

- *Conditional response probabilities*: if these probabilities do not depend on the individual covariates, they are updated through (3.16) as for the basic LM model, under the constraint of time homogeneity, that is,

  $$\phi_{y|u} = \frac{\sum_{t=1}^T \sum_{\boldsymbol{x}} \hat{a}_{u\boldsymbol{x}y}^{(t)}}{\sum_{t=1}^T \sum_{\boldsymbol{x}} \hat{b}_{u\boldsymbol{x}}^{(t)}}, \quad u = 1, \dots, k,\ y = 0, \dots, c-1.$$

  Otherwise, these parameters are updated by the iterative algorithm illustrated in Appendix 2 of Chapter 4 if these probabilities are subjected to some specific constraint. If the conditional response probabilities are modeled on the basis of the covariates, by a formulation of type (5.5), then an iterative algorithm implemented in a similar way is necessary.

- *Initial probabilities*: if these probabilities are not assumed to depend on the covariates but are unconstrained, then they are updated as in (3.17), that is,

  $$\pi_u = \frac{\sum_{\boldsymbol{x}} \hat{b}_{u\boldsymbol{x}}^{(1)}}{n}, \quad u = 1, \dots, k.$$

  Under the constraint of uniform initial probabilities, the M-step simply skips the update of these parameters, whose values are fixed as in (4.11). Finally, if the initial probabilities are assumed to depend on the individual covariates, then an iterative algorithm is necessary, which is similar to that illustrated in Appendix 2 of Chapter 4.

- *Transition probabilities*: if the transition probabilities do not depend on the covariates, then they are updated through (3.18), that is,

  $$\pi_{u|\bar{u}}^{(t)} = \frac{\sum_{\boldsymbol{x}} \hat{b}_{\bar{u}u\boldsymbol{x}}^{(t)}}{\sum_{\boldsymbol{x}} \hat{b}_{\bar{u}\boldsymbol{x}}^{(t-1)}}, \quad t = 2, \dots, T,\ \bar{u}, u = 1, \dots, k.$$

  When the Markov chain is assumed to be time homogenous and independent of the covariates, see assumption (4.12), these probabilities are updated through (4.25) or through a similar rule when the constraint of partial homogeneity is assumed. Finally, when the individual covariates are included in the latent model and affect the transition probabilities, then an iterative algorithm of the usual type is necessary; see Appendix 2 of Chapter 4.

The EM algorithm is implemented in a similar way in the multivariate case, where the main difference is in how to update the conditional response probabilities. Moreover, a strategy similar to that described in Section 4.4.1.3 can be adopted to initialize the EM algorithm even in the presence of individual covariates.

5.7 Observed information matrix, identifiability, and standard errors

To compute the information matrix for latent variable models, several methods have been proposed in the literature; see, for instance, McLachlan and Peel (2000, Section 2.15). For a method which is specific of hidden Markov models, and then may also be applied to LM models, see Lystig and Hughes (2002). In the present context, however, these methods cannot be directly applied; therefore, we prefer to use the observed information matrix, denoted by $\boldsymbol{J}(\boldsymbol{\theta})$, which is obtained by the numerical method proposed in Bartolucci and Farcomeni (2009).

According to the method above, we obtain $\boldsymbol{J}(\boldsymbol{\theta})$ as minus the numerical derivative of the score vector $\boldsymbol{s}(\boldsymbol{\theta})$. The score vector is obtained as the first derivative of the expected value of $\ell^*(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta}$ equal to the value of the parameter vector used to compute the expected frequencies $\hat{a}^{(t)}_{u\boldsymbol{x}y}$, $\hat{b}^{(t)}_{u\boldsymbol{x}}$, and $\hat{b}^{(t)}_{\bar{u}u\boldsymbol{x}}$. Note that this score vector is already used at the M-step and then computation of the observed information matrix requires a small extra code to be implemented. The observed information matrix at the maximum likelihood estimate, $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$, may be used to check local identifiability of the model and to compute the standard errors $se(\hat{\boldsymbol{\theta}})$ in the usual way.

The procedure for checking that the model is locally identifiable at $\hat{\boldsymbol{\theta}}$ consists in checking that $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is of full rank; see McHugh (1956), Rothenberg (1971), and Goodman (1974). The standard error for a parameter estimate is computed as the square root of the corresponding diagonal element of $\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}$. With reference to the models dealt with in this chapter, the validity of this procedure to obtain standard errors for $\hat{\boldsymbol{\theta}}$ was assessed by simulation in Bartolucci and Farcomeni (2009).

Obviously, the methods described above to check local identifiability and to compute the standard errors for the parameter estimates can also be used for the models illustrated in Chapters 3 and 4. However, we prefer to introduce these methods here because standard errors are typically used in connection with models including individual covariates, in order to check the significance of these covariates. On the other hand, identifiability can be checked through the multistart strategy already described in Section 3.5.1.3, suitably adapted to the case of covariates in the model.

5.8 Relaxing local independence

As already clarified in Section 5.1, the assumption of local independence may be relaxed in two different ways: by allowing conditional serial dependence or by allowing conditional contemporary dependence. In the first case, each response variable may depend on the lagged response variables even given the latent process. In the second case, which only makes sense with multivariate data, the response variables observed at the same occasion may be dependent even given the corresponding latent variable.

5.8.1 Conditional serial dependence

In the present context, relaxing local independence by allowing conditional serial dependence amounts to including the lagged response variable $Y^{(t-1)}$ into the vector of covariates $\boldsymbol{X}^{(t)}$. The augmented covariate vector is then denoted by $\boldsymbol{X}^{*(t)} = \big((\boldsymbol{X}^{(t)})' \quad Y^{(t-1)}\big)'$. Given this, all the model formulation expressed in Section 5.3 remains substantially unchanged, once we substitute $\boldsymbol{X}^{(t)}$ in each equation with $\boldsymbol{X}^{*(t)}$. Then, for instance, we write

$$\phi^{(t)}_{y|u\boldsymbol{x}^*} = f_{Y^{(t)}|U,\boldsymbol{X}^{*(t)}}(y|u,\boldsymbol{x}^*) = f_{Y^{(t)}|U,\boldsymbol{X}^{(t)},Y^{(t-1)}}(y|u,\boldsymbol{x},\bar{y}),$$

where $\boldsymbol{x}^* = \big(\boldsymbol{x}' \quad \bar{y}\big)'$, so that $\bar{y}$ is a realization of $Y^{(t-1)}$.

However, by manifest distribution of the response variables we refer to the conditional distribution of $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{X}}^*$, where $\tilde{\boldsymbol{X}}^* = \Big(\tilde{\boldsymbol{X}}' \quad Y^{(0)}\Big)'$ or $\tilde{\boldsymbol{X}}^* = \Big(\tilde{\boldsymbol{X}}' \quad (\boldsymbol{Y}^{(0)})'\Big)'$, and then we also condition on the initial responses that are assumed to be observable.

In order to illustrate the above point, we consider an example which involves univariate responses, but may be simply extended to the multivariate case. In this example, the conditional distribution of each response variable depends on the lagged response.

Example 17 — Dynamic logit model with time-varying random effects. *The logit LM model of Example 14 may be extended by assuming*

$$\log \frac{\phi^{(t)}_{1|u\boldsymbol{x}^*}}{\phi^{(t)}_{0|u\boldsymbol{x}^*}} = \alpha_u + (\boldsymbol{x}^*)'\boldsymbol{\psi}, \quad t = 1, \dots, T,\ u = 1, \dots, k,$$

so that the resulting model may be seen as an extension of the dynamic logit model (Hsiao, 2003), based on time-varying random effects. In fact, we have that

$$(\boldsymbol{x}^*)'\boldsymbol{\psi} = \boldsymbol{x}'\boldsymbol{\psi}_1 + \bar{y}\psi_2,$$

where $\boldsymbol{\psi} = \begin{pmatrix}\boldsymbol{\psi}_1' & \psi_2\end{pmatrix}'$ *and, as indicated above,* $\bar{y}$ *is a realization of the lagged response variable. In this way, the parameter* ψ_2 *measures the state dependence effect (Heckman, 1981), since it is multiplied by the realization of the lagged response variable. This model finds a natural application for the analysis of labor market data, where* $Y^{(t)}$ *is an indicator for a subject having a job position at occasion* t*. In this case,* ψ_2 *measures the effect that having a job position at a certain occasion has on the probability of having the job position at the following occasion, in addition to the effect of other observable and unobservable factors which may affect this outcome. Then, this parameter has important policy implications.*

It is easy to see that the model based on the above assumptions may be formulated through (5.5), with

$$\boldsymbol{W}_{u\boldsymbol{x}^*} = \begin{pmatrix}\boldsymbol{d}_{uk}' & (\boldsymbol{x}^*)'\end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix}(\alpha_1, \dots, \alpha_k) & \boldsymbol{\psi}'\end{pmatrix}',$$

and then in a way very similar to that in (5.9).

Since the LM model that includes the lagged response variables may be easily formulated as a model with individual covariates, we do not add details about parameter estimation. Essentially the same EM algorithm as in Section 5.6.1 may be used for this aim, whereas standard errors may be obtained as clarified in Section 5.7. The standard errors may be directly used to test if there is a form of conditional serial dependence when, as in the above example, this dependence is introduced by a single parameter. Otherwise, a likelihood ratio test between nested models needs to be used.

Concluding this section, it is worth noting that the lagged response variables may also be included among the covariates affecting the transition probabilities by a parametrization of the type illustrated in Section 5.4. A model including this component was employed by Bartolucci and

Pennoni (2007) for the analysis of capture-recapture data. Since the extension at issue is less applied, we refer the reader to this paper for further details.

5.8.2 Conditional contemporary dependence

This extension makes sense only with multivariate responses. In this case, formulating an LM model with conditional contemporary association is more complex than formulating a model with conditional serial dependence, since we need to introduce a multivariate link function. This requires a more complex notation for the conditional distribution of the response variables at a specific time occasion given the corresponding latent state.

If we admit contemporary conditional dependence, the probabilities $\phi^{(t)}_{\boldsymbol{y}|u\boldsymbol{x}} = f_{\boldsymbol{Y}^{(t)}|U^{(t)},\boldsymbol{X}^{(t)}}(\boldsymbol{y}|u,\boldsymbol{x})$ cannot be factorized in the usual way (see equation (5.4)); nevertheless, for the conditional distribution of $\boldsymbol{U}$ given $\tilde{\boldsymbol{X}}$, the conditional distribution of $\tilde{\boldsymbol{Y}}$ given $\boldsymbol{U}$ and $\tilde{\boldsymbol{X}}$, and the manifest distribution of $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{X}}$, we still have the same expressions as in (5.1), (5.2), and (5.3), respectively, with $\phi^{(t)}_{y^{(t)}|u\boldsymbol{x}}$ substituted by $\phi^{(t)}_{\boldsymbol{y}^{(t)}|u\boldsymbol{x}}$. Similarly, the usual recursion in (3.6) may be used to efficiently compute the probability mass function of the manifest distribution.

Let $\boldsymbol{p}^{(t)}_{u\boldsymbol{x}}$ denote the column vector with elements $\phi^{(t)}_{\boldsymbol{y}|u\boldsymbol{x}}$ for all possible configurations $\boldsymbol{y}$ of $\boldsymbol{Y}^{(t)}$ arranged in lexicographical order. An example of how this vector is structured is presented in the following; note that this vector has $\prod_{j=1}^{r} c_j$ elements. Then, we let $\boldsymbol{\eta}^{*(t)}_{u\boldsymbol{x}} = \boldsymbol{g}^*(\boldsymbol{p}^{(t)}_{u\boldsymbol{x}})$, where $\boldsymbol{g}^*(\cdot)$ is a *multivariate link function*, and we assume that

$$\boldsymbol{\eta}^{*(t)}_{u\boldsymbol{x}} = \boldsymbol{W}^{(t)}_{u\boldsymbol{x}}\boldsymbol{\beta}, \quad t = 1,\ldots,T,\ u = 1,\ldots,k, \tag{5.14}$$

for all possible configurations $\boldsymbol{x}$ of $\boldsymbol{X}^{(t)}$. In particular, among the possible functions of this type, we suggest adopting those including marginal logits (of the type that best suits the nature of the data, that is, local, global, etc.), marginal log-odds ratios, and similar higher-order effects. These multivariate link functions may be formulated as in Colombi and Forcina (2001) and have a structure similar to that of Glonek (1996); moreover, these link functions are strongly related to the multivariate logistic transform of Glonek and McCullagh (1995). It important to note that these link functions are easy to interpret and are simply formulated as

$$\boldsymbol{g}^*(\boldsymbol{p}^{(t)}_{u\boldsymbol{x}}) = \boldsymbol{C}^* \log(\boldsymbol{M}^*\boldsymbol{p}^{(t)}_{u\boldsymbol{x}}), \tag{5.15}$$

where $\boldsymbol{C}^*$ is a matrix of contrasts and $\boldsymbol{M}^*$ is a marginalization matrix.

A class of LM model formulated as described above, which also includes the lagged responses among the regressors, is formulated in Bartolucci and Farcomeni (2009), to which we refer the reader for a more detailed description. Here we outline this approach making explicit reference to the case of $r = 2$ response variables, where $\boldsymbol{\eta}_{u\boldsymbol{x}}^{*(t)}$ includes two sets of logits ($c_1 - 1$ for the first response variable and $c_2 - 1$ for the second response variable), in addition to a set of log-odds ratios to model the association between these variables given the latent state; the number of log-odds ratios is equal to $(c_1 - 1)(c_2 - 1)$, so that the overall dimension of $\boldsymbol{\eta}_{u\boldsymbol{x}}^{*(t)}$ is $c_1 c_2 - 1$. How to construct the matrices $\boldsymbol{C}^*$ and $\boldsymbol{M}^*$, depending on the type of logit, is described in Appendix 1, where we also show how to compute the inverse of $\boldsymbol{g}^*(\cdot)$, that is, how to obtain $\boldsymbol{p}_{u\boldsymbol{x}}^{(t)}$ on the basis of a given value of $\boldsymbol{\eta}_{u\boldsymbol{x}}^{*(t)}$.

The following example helps to understand how to formulate an LM model as indicated above for the case of two response variables. For the notation, we still use $\phi_{jy|u\boldsymbol{x}}^{(t)}$ for the conditional probability that $Y_j^{(t)} = y$, given $U^{(t)} = u$ and $\boldsymbol{X}^{(t)} = \boldsymbol{x}$, that is,

$$\phi_{jy|u\boldsymbol{x}}^{(t)} = f_{Y_j^{(t)}|U^{(t)},\boldsymbol{X}^{(t)}}(y|u,\boldsymbol{x}),$$

which has a different meaning with respect to

$$\phi_{(y_1,y_2)|u\boldsymbol{x}}^{(t)} = f_{Y_1^{(t)},Y_2^{(t)}|U^{(t)},\boldsymbol{X}^{(t)}}(y_1,y_2|u,\boldsymbol{x});$$

the latter is a more explicit way of writing $\phi_{\boldsymbol{y}|u\boldsymbol{x}}^{(t)}$ considering that in the bivariate case $\boldsymbol{y}$ has two elements, y_1 and y_2.

Example 18 — Marginal model for bivariate responses. *Consider the case of $r = 2$ response variables with $c_1 = 2$ and $c_2 > 2$ levels, which are treated with logits of type local and global, respectively. Overall, there are c_2 logits. About the logit for the first variable, we assume*

$$\log \frac{\phi_{11|u\boldsymbol{x}}^{(t)}}{\phi_{10|u\boldsymbol{x}}^{(t)}} = \alpha_{1u} + \boldsymbol{x}'\boldsymbol{\psi}_1$$

as in Example 14. About the two global logits of the second variable, we assume

$$\log \frac{\phi_{2y|u\boldsymbol{x}}^{(t)} + \cdots + \phi_{2,c_2-1|u\boldsymbol{x}}^{(t)}}{\phi_{20|u\boldsymbol{x}}^{(t)} + \cdots + \phi_{2,y-1|u\boldsymbol{x}}^{(t)}} = \alpha_{2uy} + \boldsymbol{x}'\boldsymbol{\psi}_2, \quad y = 1, \ldots, c_2 - 1,$$

as in Example 15. Finally, we have $c_2 - 1$ log-odds ratios parametrized *as follows:*

$$\log \frac{(\phi^{(t)}_{(0,0)|u\boldsymbol{x}} + \cdots + \phi^{(t)}_{(0,y-1)|u\boldsymbol{x}})(\phi^{(t)}_{(1,y)|u\boldsymbol{x}} + \cdots + \phi^{(t)}_{(1,c_2-1)|u\boldsymbol{x}})}{(\phi^{(t)}_{(0,y)|u\boldsymbol{x}} + \cdots + \phi^{(t)}_{(0,c_2-1)|u\boldsymbol{x}})(\phi^{(t)}_{(1,0)|u\boldsymbol{x}} + \cdots + \phi^{(t)}_{(1,y-1)|u\boldsymbol{x}})} = \psi_{3y},$$

for $y = 1, \ldots, c_2 - 1$*. In this way, the strength of the association depends neither on the covariates nor on the latent state.*

In order to express how to construct the matrices $\boldsymbol{C}^*$ *and* $\boldsymbol{M}^*$ *in this case, suppose that* $c_2 = 3$*; then the second response variable may be equal to 0, 1, or 2 and the vector* $\boldsymbol{p}^{(t)}_{u\boldsymbol{x}}$ *has the following structure:*

$$\boldsymbol{p}^{(t)}_{u\boldsymbol{x}} = \begin{pmatrix} \phi^{(t)}_{(0,0)|u\boldsymbol{x}} \\ \phi^{(t)}_{(0,1)|u\boldsymbol{x}} \\ \phi^{(t)}_{(0,2)|u\boldsymbol{x}} \\ \phi^{(t)}_{(1,0)|u\boldsymbol{x}} \\ \phi^{(t)}_{(1,1)|u\boldsymbol{x}} \\ \phi^{(t)}_{(1,2)|u\boldsymbol{x}} \end{pmatrix}. \tag{5.16}$$

Therefore, we have

$$\boldsymbol{C}^* = \left(\begin{array}{cccccccccccccc} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 \end{array}\right)$$

and then

$$\boldsymbol{M}^* = \left(\begin{array}{cccccc} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right).$$

Finally, we have $\boldsymbol{\beta} = ((\alpha_{11}, \alpha_{211}, \alpha_{212}, \ldots, \alpha_{1k}, \alpha_{2k1}, \alpha_{2k2})\, \boldsymbol{\psi}'_1\, \boldsymbol{\psi}'_2\, (\psi_{31}, \psi_{32}))'$

and

$$\boldsymbol{W}_{u\boldsymbol{x}}^{(t)} = \left(\begin{array}{ccccc} \boldsymbol{d}_{uk}' \otimes \boldsymbol{d}_{13}' & \boldsymbol{x}' & \boldsymbol{0}' & 0 & 0 \\ \hline \boldsymbol{d}_{uk}' \otimes \boldsymbol{d}_{23}' & \boldsymbol{0}' & \boldsymbol{x}' & 0 & 0 \\ \boldsymbol{d}_{uk}' \otimes \boldsymbol{d}_{33}' & \boldsymbol{0}' & \boldsymbol{x}' & 0 & 0 \\ \hline \boldsymbol{0}' & \boldsymbol{0}' & \boldsymbol{0}' & 1 & 0 \\ \boldsymbol{0}' & \boldsymbol{0}' & \boldsymbol{0}' & 0 & 1 \end{array}\right).$$

In order to estimate the LM models based on parametrization (5.14), we can use an EM algorithm having the usual structure, even if the M-step of this algorithm is more complex to implement. However, since we can express the marginal link function as in (5.15), the complexity is not so high. In any case, we refer the reader to Bartolucci and Farcomeni (2009) for a detailed description of the estimation algorithm, in a version that also allows us to include serial conditional dependence in different forms.

5.9 Higher order extensions

The LM model, in all of the formulations illustrated up to this point, is based on the assumption that the underlying Markov chain is of first order. In certain contexts this assumption may be restrictive. In the following, we show, in particular, how to extend the model so that the underlying Markov chain is of second-order. LM models of this type have already been considered by Wiggins (1973). The main point is that, as shown by du Preez (1997) and Bartolucci and Solis-Trapala (2010), the second-order model may be reformulated as a first-order model with a particular state-space structure and suitable constraints on the initial and transition probabilities; then, maximum likelihood estimation may be carried out by an EM algorithm having the same structure as the one used up to this point.

Suppose, for simplicity, that individual covariates are not included in the model and let u'' denote a generic latent state of the second-order LM model, with $u'' = 1, \dots, k''$. Then, this model is based on the initial probabilities

$$\pi''_{u''} = f_{U^{(1)}}(u''), \quad u'' = 1, \dots, k''.$$

Moreover, for the second time occasion, we have the transition probabilities

$$\pi''^{(2)}_{u''|\bar{u}''} = f_{U^{(2)}|U^{(1)}}(u''|\bar{u}''), \quad \bar{u}'', u'' = 1, \dots, k'',$$

whereas, for $t = 3, \dots, T$, we introduce the transition probabilities

$$\pi''^{(t)}_{u''|\bar{\bar{u}}''\bar{u}''} = f_{U^{(t)}|U^{(t-2)},U^{(t-1)}}(u''|\bar{\bar{u}}'', \bar{u}''), \quad \bar{\bar{u}}'', \bar{u}'', u'' = 1, \dots, k''.$$

In the previous expressions, we follow the general notation, and then by $\bar{u}''$ we denote the latent state at lag 1 and by $\bar{\bar{u}}''$ that at lag 2. Even these probabilities may be parameterized in a suitable way, on the basis of available covariates, or appropriately constrained in order to reduce the number of model parameters. The other model components, and in particular the conditional response probabilities, are left unchanged; for these probabilities, we use the notation $\phi''_{y|u''}$ or $\phi''_{jy|u''}$, which have the usual meaning when the model is in the basic version. In the univariate case, the number of free parameters is then equal to

$$\begin{aligned} \#\text{par} \quad = \quad & \underbrace{k''(c-1)}_{\phi''_{y|u''}} + \underbrace{k''-1}_{\pi''_{u''}} + \underbrace{k''(k''-1)}_{\pi''^{(2)}_{u''|\bar{u}''}} \\ & + \underbrace{(T-2)(k'')^2(k''-1)}_{\pi''^{(t)}_{u''|\bar{\bar{u}}''\bar{u}'}}. \end{aligned} \tag{5.17}$$

A second-order LM model, based on the above parameters, may be expressed as a first-order model with an augmented state space having $k = (k'')^2$ states. Each state u of the first-order model corresponds to a pair of states $(\bar{u}''(u), u''(u))$ of the second-order model. It is important to note that transition between two states, $\bar{u}$ and u, of the augmented states space is only possible if $u''(\bar{u}) = \bar{u}''(u)$. Then, we denote by $\mathcal{S}_{\bar{u}}$ the set of states which are compatible with $\bar{u}$, that is, the states on which it is possible to move starting from $\bar{u}$; this set has the following structure:

$$\mathcal{S}_{\bar{u}} = \{u: \ u = 1, \dots, k, \ u''(\bar{u}) = \bar{u}''(u)\}.$$

Also let $\bar{\mathcal{S}}_{\bar{u}}$ denote the complement of $\mathcal{S}_{\bar{u}}$ with respect to the state space $\{1, \dots, k\}$. This structure of the state space implies certain constraints that will be shown in the following.

In order to clarify the above notation, consider the following example.

Example 19 — Equivalence between first-order and second-order LM models (part 1). *Suppose that the second-order LM model is based on $k'' = 3$ states. Then, the equivalent first-order model has $k = 9$ latent states that have the meaning showed in Table 5.1.*

Consequently, we have that

$$\mathcal{S}_1 = \{1, 2, 3\}, \quad \bar{\mathcal{S}}_1 = \{4, 5, 6, 7, 8, 9\}$$

TABLE 5.1
Equivalence between latent states

u	$(\bar{u}''(u), u''(u))$
1	$(1,1)$
2	$(1,2)$
3	$(1,3)$
4	$(2,1)$
5	$(2,2)$
6	$(2,3)$
7	$(3,1)$
8	$(3,2)$
9	$(3,3)$

since it is possible to move from state 1 to state 1, 2, or 3 for the first-order LM model. Similarly, we have

$$\mathcal{S}_2 = \{4,5,6\}, \quad \bar{\mathcal{S}}_2 = \{1,2,3,7,8,9\}$$

and so on for $\mathcal{S}_3,\dots,\mathcal{S}_9$ *and their complementary sets.*

The first-order LM model, which is equivalent to the second-order model, has initial probabilities

$$\pi_u = \pi''_{u''(u)}, \quad u = 1,\dots,k, \tag{5.18}$$

where, because different values of u'' correspond to the same value of u, the same parameter $\pi''_{u''}$ is replicated k times. Similarly, when $t = 2$, its transition probabilities are

$$\pi^{(t)}_{u|\bar{u}} = \begin{cases} \pi''^{(2)}_{u''(u)|\bar{u}''(u)}, & u \in \mathcal{S}_{\bar{u}}, \\ 0, & u \in \bar{\mathcal{S}}_{\bar{u}}, \end{cases} \tag{5.19}$$

where $(\bar{u}''(u), u''(u))$ depends on u; again, there are replicates among these transition probabilities. Moreover, for $t = 3,\dots,T$, we have

$$\pi^{(t)}_{u|\bar{u}} = \begin{cases} \pi''^{(t)}_{u''(u)|\bar{u}''(u)\bar{u}''(\bar{u})}, & u \in \mathcal{S}_{\bar{u}}, \\ 0, & u \in \bar{\mathcal{S}}_{\bar{u}}. \end{cases} \tag{5.20}$$

Finally, we have the conditional response probabilities

$$\phi_{y|u} = \phi''_{y|u''(u)}, \quad y = 0,\dots,c-1,\ u = 1,\dots,k, \tag{5.21}$$

where the same conditional response probability is replicated k times in the first-order model.

The parametrization of initial and transition probabilities of the first-order model is clarified by the following example.

Example 20 — Equivalence between first-order and second-order LM models (part 2). *In the same context as in Example 19, the vector of initial probabilities of the first-order model has the following structure*

$$(\pi_1'', \pi_2'', \pi_3'', \pi_1'', \pi_2'', \pi_3'', \pi_1'', \pi_2'', \pi_3'')',$$

whereas the transition matrix for $t = 2$*, based on (5.19), has the following structure:*

$$\begin{pmatrix}
\pi_{1|1}^{\prime\prime(2)} & \pi_{2|1}^{\prime\prime(2)} & \pi_{3|1}^{\prime\prime(2)} & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & \pi_{1|2}^{\prime\prime(2)} & \pi_{2|2}^{\prime\prime(2)} & \pi_{3|2}^{\prime\prime(2)} & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & \pi_{1|3}^{\prime\prime(2)} & \pi_{2|3}^{\prime\prime(2)} & \pi_{3|3}^{\prime\prime(2)}\\
\pi_{1|1}^{\prime\prime(2)} & \pi_{2|1}^{\prime\prime(2)} & \pi_{3|1}^{\prime\prime(2)} & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & \pi_{1|2}^{\prime\prime(2)} & \pi_{2|2}^{\prime\prime(2)} & \pi_{3|2}^{\prime\prime(2)} & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & \pi_{1|3}^{\prime\prime(2)} & \pi_{2|3}^{\prime\prime(2)} & \pi_{3|3}^{\prime\prime(2)}\\
\pi_{1|1}^{\prime\prime(2)} & \pi_{2|1}^{\prime\prime(2)} & \pi_{3|1}^{\prime\prime(2)} & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & \pi_{1|2}^{\prime\prime(2)} & \pi_{2|2}^{\prime\prime(2)} & \pi_{3|2}^{\prime\prime(2)} & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & \pi_{1|3}^{\prime\prime(2)} & \pi_{2|3}^{\prime\prime(2)} & \pi_{3|3}^{\prime\prime(2)}
\end{pmatrix}$$

Finally, for $t = 3, \ldots, T$*, assumption (5.20) implies that we have the following transition matrices:*

$$\begin{pmatrix}
\pi_{1|11}^{\prime\prime(t)} & \pi_{2|11}^{\prime\prime(t)} & \pi_{3|11}^{\prime\prime(t)} & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & \pi_{1|12}^{\prime\prime(t)} & \pi_{2|12}^{\prime\prime(t)} & \pi_{3|12}^{\prime\prime(t)} & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & \pi_{1|13}^{\prime\prime(t)} & \pi_{2|13}^{\prime\prime(t)} & \pi_{3|13}^{\prime\prime(t)}\\
\pi_{1|21}^{\prime\prime(t)} & \pi_{2|21}^{\prime\prime(t)} & \pi_{3|21}^{\prime\prime(t)} & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & \pi_{1|22}^{\prime\prime(t)} & \pi_{2|22}^{\prime\prime(t)} & \pi_{3|22}^{\prime\prime(t)} & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & \pi_{1|23}^{\prime\prime(t)} & \pi_{2|23}^{\prime\prime(t)} & \pi_{3|23}^{\prime\prime(t)}\\
\pi_{1|31}^{\prime\prime(t)} & \pi_{2|31}^{\prime\prime(t)} & \pi_{3|31}^{\prime\prime(t)} & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & \pi_{1|32}^{\prime\prime(t)} & \pi_{2|32}^{\prime\prime(t)} & \pi_{3|32}^{\prime\prime(t)} & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & \pi_{1|33}^{\prime\prime(t)} & \pi_{2|33}^{\prime\prime(t)} & \pi_{3|33}^{\prime\prime(t)}
\end{pmatrix}.$$

The first-order LM model may be estimated by an EM algorithm taking into account the above constraints. Then, its parameter estimates must be expressed in terms of the parameters of the initial model on the basis of (5.18), (5.19), (5.20), and (5.21).

An important point is how to compare the first-order with the second-order LM model. This comparison may be carried out by a likelihood ratio test that has null asymptotic distribution of χ^2-type. The number of degrees of freedom of this distribution is equal to the number of constraints required to reduce a second-order model to a first-order model with the same number k'' of latent states. The constraints at issue are

$$\pi''^{(t)}_{u|1\bar{u}} = \cdots = \pi''^{(t)}_{u|k''\bar{u}}, \quad t = 3,\dots,T,\ \bar{\bar{u}},\bar{u},u = 1,\dots,k''.$$

The number of independent constraints are then $(T-2)k''(k''-1)^2$, which corresponds to the difference in the number of parameters between a second-order LM model, see equation (5.17), and a first-order LM model, see equation (3.1).

5.10 Applications

In the following, we illustrate the LM models with covariates by the analysis of two of the datasets introduced in Chapter 1 carried out in Bartolucci et al. (2007) and in Bartolucci and Farcomeni (2009).

5.10.1 Criminal conviction history dataset

For the analysis of these data, Bartolucci et al. (2007) adopted a model which represents an extension of the multivariate LM model we selected in Section 4.6.2. The extension essentially consists in the inclusion of the covariate *gender*.

As usual, the first step of the analysis was the choice of the number of latent states. For this aim, Bartolucci et al. (2007) fitted, for an increasing number of latent states between 1 and 6, a multivariate LM model based on the following assumptions:

- the conditional response probabilities are time homogenous and do not depend on the covariate; moreover, the probability of committing a crime is constrained to 0 for the first latent state;
- the initial probabilities are distinct for males and females;
- the transition probabilities are time heterogeneous and distinct for males and females.

The results of this preliminary fitting are reported in Table 5.2 and lead to selecting $k = 5$ latent states. The model with this number of states, denoted hereafter by M_9, has maximum log-likelihood of -107871, with 248 parameters, and then $BIC = 218439$.

Then, Bartolucci et al. (2007) tried to simplify model M_8 by suitable

TABLE 5.2
Results of a preliminary fitting to choose the number of latent states

k	$\hat{\ell}$	#par	BIC
1	−166294	10	332696
2	−114297	32	228941
3	−110072	84	221058
4	−108694	156	219084
5	−107871	248	218439
6	−107399	360	218713

constraints on the initial and transition probabilities. In particular, for the model with the same transition probabilities for males and females, but with different initial probabilities, we have $BIC = 218233$ with 148 parameters; this model, denoted by M_{10}, is therefore preferable to model M_9. According to model M_{10}, males and females may differ in terms of the pattern of criminal activity since the probability of belonging to a particular latent state in the first age band may be different. However, they do not have a different way of moving between the latent states.

The next step for finding a suitable model for the criminal conviction history dataset was based on trying different types of partial time homogeneity of the latent Markov chain, as formulated in assumption (4.13). We recall that, according to this assumption, until time occasion T^* we have different transition probabilities than for the following occasions. In this regard, Table 5.3 shows the value of BIC for all the possible values of T^*.

The lowest value of the BIC index is 217896 for $T^* = 2$, which is lower than the value obtained for the heterogeneous model. The model selected in this way, hereafter denoted by M_{11}, assumes that, apart from the transition probabilities from the first to the second occasion (from age band 10–15 to 16–20), all the other transition probabilities are time homo-

TABLE 5.3
Results from fitting a constrained version of model M_{10} which includes partial time homogeneity for different values of T^*

T^*	$\hat{\ell}$	#par	BIC
2	−108470	88	217896
3	−108533	88	218023
4	−108541	88	218039
5	−108619	88	218195

TABLE 5.4
Estimates of the conditional probabilities of conviction, $\phi_{j1|u}$, under model M_{11}

	$\hat{\phi}_{j1\|u}$				
j	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$
1	0.000	0.003	0.158	0.018	0.227
2	0.000	0.003	0.029	0.003	0.026
3	0.000	0.032	0.006	0.016	0.487
4	0.000	0.000	0.005	0.002	0.039
5	0.000	0.096	0.067	0.546	0.777
6	0.000	0.000	0.019	0.130	0.149
7	0.000	0.016	0.091	0.010	0.233
8	0.000	0.000	0.075	0.016	0.099
9	0.000	0.000	0.005	0.003	0.044
10	0.000	0.000	0.060	0.039	0.347

TABLE 5.5
Estimated initial probabilities under model M_{11}

u	$\hat{\pi}_{u\|0}$	$\hat{\pi}_{u\|1}$
1	0.496	0.963
2	0.472	0.020
3	0.000	0.000
4	0.000	0.016
5	0.033	0.000

Note: $\pi_{u|0}$ are the initial probability for males and $\pi_{u|1}$ are those for females

geneous. The number of parameters of this model is 88 since it retains different initial and equal transition probabilities for males and females and time-homogenous conditional response probabilities. We also recall that, with reference to the first state, the probability of committing any type of crime is constrained to 0.

In order to illustrate the results of the selected model, M_{11}, in Table 5.4 we report the estimated conditional probabilities of conviction for each offense group and latent state. Moreover, in Table 5.5 we show the estimated initial probabilities for males and females, and in Table 5.6 we show the common estimated transition probabilities from age band 10–15 to age band 16–20, whereas Table 5.7 contains the estimated transition probabilities from one age band to the next one for offenders aged 16 and over.

Based on the results in Table 5.5, Bartolucci et al. (2007) character-

TABLE 5.6
Estimates of the transition probabilities $\pi^{*(1)}_{u|\bar{u}}$ from age band 10–15 to age band 16–20, obtained under model M_{11} for both males and females

	$\hat{\pi}^{*(1)}_{u\|\bar{u}}$				
$\bar{u}$	$u=1$	$u=2$	$u=3$	$u=4$	$u=5$
1	0.960	0.009	0.003	0.028	0.000
2	0.068	0.648	0.140	0.053	0.092
3	1.000	0.000	0.000	0.000	0.000
4	0.846	0.000	0.005	0.149	0.000
5	0.040	0.175	0.033	0.000	0.753

TABLE 5.7
Estimates of the transition probabilities $\pi^{*(2)}_{u|\bar{u}}$ from one age band to another for male and female offenders over 16, obtained under model M_{11}

	$\hat{\pi}^{*(2)}_{u\|\bar{u}}$				
$\bar{u}$	$u=1$	$u=2$	$u=3$	$u=4$	$u=5$
1	0.993	0.000	0.005	0.012	0.000
2	0.327	0.371	0.206	0.078	0.018
3	0.583	0.000	0.395	0.015	0.007
4	0.809	0.001	0.008	0.180	0.002
5	0.000	0.276	0.217	0.039	0.468

ized the latent states as those of “nonoffenders” (1), “incidental offender” (2), “violent offenders” (3), “theft and fraud offenders” (4), and “high frequency and varied offenders” (5). Then, considering also the estimates in Table 5.5, we conclude that, at the beginning of the period of observation (age band 10–15), the probability of being a nonoffender is much higher for a female than for a male. Moreover, on the basis of the transition probabilities in Tables 5.6 and 5.7, we conclude that these subjects have a high probability of persistence in the same latent state, and then a very low probability of becoming offenders, as they become older. On the basis of the estimated transition probabilities, we can also study how the behavior of a subject evolves over time. For instance, we observe that violent offenders and theft and fraud offenders have a high probability of dropping out of crime. However, for violent offenders, the probability of this drop out is much higher when they are young (transition from 10–15 to 16–20) than when they are older (aged sixteen and over). A significant difference between the transition probabilities is not observed for theft and fraud offenders.

TABLE 5.8
Log-likelihood, number of parameters, and AIC and BIC indices obtained from fitting the model with 1 to 5 latent states

k	log-lik.	#par.	AIC	BIC
1	−6219.0	37	12512	12707
2	−6050.0	44	12188	12420
3	−6011.5	53	12129	12409
4	−6004.7	64	12137	12475
5	−5993.6	77	12141	12548

5.10.2 Labor market dataset

In this section, we outline the analysis of the dataset described in Section 1.4.3 that was carried out in Bartolucci and Farcomeni (2009) by an LM model with covariates.

The analysis is focused on studying the dependence between the response variables *fertility* and *employment*, taking into account the observable covariates and a time-varying unobserved heterogeneity effect between subjects. For this aim, Bartolucci and Farcomeni (2009) used a model formulated as described in Section 5.8.2, see in particular Example 18, which also allows for serial dependence between the response variables and is based on an LM model which has initial probabilities depending on the covariates and time-homogenous transition probabilities which are independent of the covariates.

Following the usual approach, Bartolucci and Farcomeni (2009) initially fitted the model at issue for an increasing number of latent states. The results of this fitting are shown in Table 5.8, which has a structure similar to that of the tables reported for the previous applications.

On the basis of the results in Table 5.8, we conclude that a suitable number of latent states is $k = 3$, because it corresponds to the minimum value of both AIC and BIC; we denote the model with $k = 3$ by M_1.

The estimates of the parameters affecting the conditional distribution of the response variables given the latent process are reported in Table 5.9 under model M_1 and under versions of this model based on a different number of latent states. In order to properly interpret the results in this table, we have to consider that the adopted model is based on two logits for the conditional distribution of each response variable given the covariates, the lagged responses, and the latent state. Moreover, the strength of the contemporary dependence between these variables is measured by one conditional log-odds ratio, which is assumed to be constant with respect to the covariates and the latent variables. Consequently, Table 5.9 is divided into three panels corresponding to the

TABLE 5.9
Estimates of the regression parameters affecting the conditional distribution of the response variables given the latent process

	k				
Effect	1	2	3	4	5
			logit fertility		
intercept*	−1.807	−2.072	−2.117	−2.198	−2.101
race	−0.230**	−0.230**	−0.235**	−0.243**	−0.239**
age†	−0.216**	−0.218**	−0.223**	−0.226**	−0.224**
(age†)^2/100	−1.112**	−1.122**	−1.135**	−1.153**	−1.107**
education†	0.152**	0.154**	0.160**	0.162**	0.160**
child 1–2	0.183**	0.187**	0.177**	0.177**	0.170**
child 3–5	−0.360**	−0.374**	−0.389**	−0.390**	−0.388**
child 6–13	−0.594**	−0.605**	−0.611**	−0.613**	−0.608**
child 14–	−0.879**	−0.885**	−0.893**	−0.897**	−0.903**
income†/1000	0.002	0.002	0.002	0.002	0.002
lag fertility	−1.476**	−1.469**	−1.482**	−1.452**	−1.499**
lag employment	−0.163	0.212	0.444**	0.443**	0.427**
			logit employment		
intercept*	−0.688	0.523	−0.010	−0.205	0.087
race	0.099	0.125	0.134	0.163	0.192
age†	0.015**	0.028	0.068**	0.070**	0.074**
(age†)^2/100	−0.103	−0.093	0.045	0.109	−0.205
education†	0.102**	0.125	0.096**	0.104**	0.121**
child 1–2	−0.116**	−0.174	−0.089	−0.010	−0.031
child 3–5	−0.234**	−0.219	−0.190**	−0.1613	−0.146
child 6–13	−0.062	0.012	−0.006	0.030	0.034
child 14–	−0.010	0.052	0.065	0.086	0.160
income†/1000	−0.009**	−0.009	−0.013**	−0.013**	−0.014**
lag fertility	−0.478**	−0.733**	−0.704**	−0.654**	−0.747**
lag employment	2.949**	1.571**	1.008**	1.0789**	0.746**
log-odds ratio	−1.213**	−1.286**	−1.130**	−1.651**	−1.173**

Note: *average of the support points based on posterior probabilities
†minus the sample average
**significant at the 5% level

regression coefficients for the logit of fertility, the regression coefficients for the logit of employment, and the conditional log-odds ratio.

On the basis of the above estimates we conclude that, under model M_1, race has a significant effect on fertility, but not on employment. Moreover, age has a stronger effect on fertility than on employment, education has a significant effect on both fertility and employment, whereas the number of children in the family strongly affects only the first response variable and income of the husband strongly affects only the second one.

The log-odds ratio between the two response variables, given the latent state, is negative and highly significant, meaning that there is a negative dependence between the two response variables, even given the covariates and the latent variables. On the other hand, lagged fertility has a significant negative effect on both response variables and lagged employment has a significant effect, which is positive, on both response variables. Therefore, Bartolucci and Farcomeni (2009) concluded that fertility has a negative effect on the probability of having a job position in the same year of the birth and the following one. Employment is positively serially correlated (as a consequence of the state dependence effect) and fertility is negatively serially correlated.

Regarding the distribution of the latent process, in Tables 5.10 and 5.11 we report the estimates of the support points corresponding to each latent state, the estimates of the parameters of the model on the initial probabilities of the latent states, and the estimates of the transition probabilities.

The results in Table 5.10 lead to the conclusion that the first latent state corresponds to women with the highest propensity to fertility and the lowest propensity to have a job position. We are referring to the "residual" propensity, that is, the propensity that cannot be explained on the basis of the observable covariates. On the contrary, the third latent state corresponds to subjects with the lowest propensity to fertility and the highest propensity to have a job position. Finally, the second state is associated with intermediate levels of both propensities. The two propensities are negatively correlated.

On the basis of the above estimates, Bartolucci and Farcomeni (2009) derived that 78.5% of the women started and persisted in the same latent state for the entire period, whereas for 21.5% of the women we have one

TABLE 5.10
Estimated support points for each latent state and estimated parameters of the model for the initial probabilities

Latent	Support points		Initial prob.		
state (u)	Fert.	Empl.	Interc.	Fert.	Empl.
1	−1.35	−5.36	–	–	–
2	−1.86	−1.07	0.78	0.34	0.86
3	−2.51	2.21	0.37	0.02	4.25**

Note: **significant at the 5% level

TABLE 5.11
Estimates of the transition probabilities $\pi_{u|\bar{u}}$

	$\hat{\pi}_{u\|\bar{u}}$		
$\bar{u}$	$u=1$	$u=2$	$u=3$
1	0.95	0.05	0.00
2	0.07	0.89	0.04
3	0.00	0.09	0.91

or more transition between states. Moreover, the hypothesis that the transition is diagonal is rejected; this means that the LM model, which admits time-varying unobserved heterogeneity, is preferable to its latent class version, which assumes time constancy of the latent variables.

Appendix 1: Multivariate link function

The multivariate logistic link function $\boldsymbol{g}^*(\cdot)$ in (5.15) is based on marginal (with respect to the other response variable) logits and the log-odds ratios of type *reference category*, *local*, *global*, *continuation*, or *reverse continuation*. Log-odds ratios are defined as contrasts between conditional logits, and their definition depends on the type of logit chosen for each response variable.

Provided that the elements in $\boldsymbol{p}_{\boldsymbol{ux}}^{(t)}$ are arranged with the response configurations in lexicographical order, as in (5.16), and following Colombi and Forcina (2001), the link function may be written as in (5.15) with $\boldsymbol{C}^*$ and $\boldsymbol{M}^*$ defined as follows:

$$\boldsymbol{C}^* = \begin{pmatrix} \boldsymbol{C}_1 & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{C}_2 & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{C}_1 \otimes \boldsymbol{C}_2 \end{pmatrix}, \quad \boldsymbol{M}^* = \begin{pmatrix} \boldsymbol{M}_1 \otimes \boldsymbol{1}'_{C_2} \\ \boldsymbol{1}'_{C_1} \otimes \boldsymbol{M}_2 \\ \boldsymbol{M}_1 \otimes \boldsymbol{M}_2 \end{pmatrix},$$

where $\boldsymbol{C}_1$ is the matrix of contrasts and $\boldsymbol{M}_1$ is the marginalization matrix defined according to the type of logit used for the first response variable and the number of its categories (c_1), as described in Appendix 1 of Chapter 4; the matrices $\boldsymbol{C}_2$ and $\boldsymbol{M}_2$ are defined in a similar way for the second variable.

In order to compute the inverse of $\boldsymbol{g}^*(\cdot)$, that is, to compute the probability vector $\boldsymbol{p}_{\boldsymbol{ux}}^{(t)}$ on the basis of a given value $\bar{\boldsymbol{\eta}}_{\boldsymbol{ux}}^{*(t)}$ of $\boldsymbol{\eta}_{\boldsymbol{ux}}^{*(t)}$, we use, for each type of logit, a Newton algorithm. In order to take into account that the elements of this probability vector have to sum to 1, the Newton algorithm uses the representation

$$\boldsymbol{p}_{u\boldsymbol{x}}^{(t)} = \frac{1}{\boldsymbol{1}' \exp(\boldsymbol{G}^* \boldsymbol{\xi}_{u\boldsymbol{x}}^{(t)})} \exp(\boldsymbol{G}^* \boldsymbol{\xi}_{u\boldsymbol{x}}^{(t)}), \quad \boldsymbol{G}^* = \begin{pmatrix} \boldsymbol{0}'_{c_1c_2-1} \\ \boldsymbol{I}_{c_1c_2-1} \end{pmatrix},$$

where the vector of log-linear parameters $\boldsymbol{\xi}_{u\boldsymbol{x}}^{(t)}$ belongs to $\mathbb{R}^{c_1c_2-1}$. Then, starting from an initial value of this vector of parameters, the algorithm iteratively updates it by adding

$$\boldsymbol{R}_{u\boldsymbol{x}}^{*(t)} (\bar{\boldsymbol{\eta}}_{u\boldsymbol{x}}^{*(t)} - \boldsymbol{\eta}_{u\boldsymbol{x}}^{*(t)})$$

until convergence, where

$$\boldsymbol{R}_{u\boldsymbol{x}}^{*(t)} = \frac{\partial \boldsymbol{\xi}_{u\boldsymbol{x}}^{(t)}}{\partial (\boldsymbol{\eta}_{u\boldsymbol{x}}^{*(t)})'} = [\boldsymbol{C}^* (\boldsymbol{M}^* \boldsymbol{p}_{u\boldsymbol{x}}^{(t)})^{-1} \boldsymbol{M}^* \mathrm{diag}(\boldsymbol{p}_{u\boldsymbol{x}}^{(t)}) \boldsymbol{G}^*]^{-1}.$$

6

Including random effects and extension to multilevel data

6.1 Introduction

In this chapter, we consider the extension of the latent Markov (LM) approach to include random effects. These effects may be assumed to have a continuous or a discrete distribution and, as the individual covariates (see Chapter 5), they may be included in the measurement model or in the latent model. In the first case we account for another form of heterogeneity, with respect to that captured by the Markov chain, which is time invariant. In the second case we allow initial and transition probabilities of the latent Markov chain to differ between sample units in a way that does not depend on only the observable covariates.

The inclusion of random effects is also important to account for data having a hierarchical or multilevel structure. In this case, random effects are used to model the influence of each cluster on the responses provided by the subjects who are included in the cluster. A typical example is when we consider responses to test items of school children who are grouped in classes.

In the following, after a description of the model assumptions necessary to include random effects in an LM model, we illustrate parameter estimation by an extension of the Expectation-Maximization (EM) algorithm for LM models with covariates. In order to keep the presentation simple, we discuss only the case of univariate data and that of random effects having a discrete distribution. An application based on education data is also illustrated.

6.2 Random-effects formulation

Typically, random effects have the role of representing the effect of unobservable covariates, in addition to the effect of observable covariates, on the responses provided by the sample units. Therefore, we are in a context similar to that of Chapter 5, where we already clarified that observable covariates may be included in the measurement model, and then affect the conditional distribution of the response variables given the latent process, or in the latent model, and then they affect the initial and transition probabilities. We also clarified that, usually, observable covariates are included alternatively in the measurement model or in the latent model mainly in order to avoid interpretability problems.

For the random effects we adopt the same criterion as above, and as we show below, this reflects on the basic notation that is an extension of the one illustrated in Section 5.2.

6.2.1 Model assumptions

First of all, in order to distinguish the random effects from the latent Markov process, we represent the first by the unobservable variable U_1. It is important to stress that these random effects are time invariant. Moreover, the latent process is now denoted by $\boldsymbol{U}_2 = (U_2^{(1)}, \ldots, U_2^{(T)})$, which substitutes the symbol $\boldsymbol{U}$ used in the previous chapters. The notation for the response variables and the covariates remains unchanged and then by $Y^{(t)}$ we denote the response at occasion t and by $\boldsymbol{X}^{(t)}$ the corresponding vector of covariates, where $t = 1, \ldots, T$. The vector of all responses is also denoted by $\tilde{\boldsymbol{Y}}$ and the vector of all covariates is denoted by $\tilde{\boldsymbol{X}}$.

As mentioned in Section 6.1, the random effects may have a continuous distribution (typically a normal distribution) or a discrete distribution. The main reference for the first type of approach is Altman (2007), whereas, for the second approach, the main reference is Maruotti (2011). In order to simplify the presentation, we explicitly consider the case of discrete random effects. In practice, in this case U_1 corresponds to an additional discrete latent variable, with respect to those included in the LM models previously illustrated, which has k_1 support points (corresponding to latent classes) and corresponding mass probabilities denoted by ω_{u_1}, $u_1 = 1, \ldots, k_1$. Accordingly, we denote by k_2 the number of latent states, that corresponds to the number of support points of every latent variable $U_2^{(t)}$.

6.2.1.1 Random effects in the measurement model

When the random effects are included in the measurement model, the basic assumption is that the response variables in $\tilde{\boldsymbol{Y}}$ are conditionally independent given the latent variable U_1, the latent variables in $\boldsymbol{U}_2$, and the covariates in $\tilde{\boldsymbol{X}}$. Another basic assumption is that of independence between U_1 and $\boldsymbol{U}_2$ and that every response variable $Y^{(t)}$ depends only on U_1 and the time-specific variables $U_2^{(t)}$ and $\boldsymbol{X}^{(t)}$. Then, we introduce the notation

$$\phi_{y|u_1u_2\boldsymbol{x}}^{(t)} = f_{Y^{(t)}|U_1,U_2^{(t)},\boldsymbol{X}^{(t)}}(y|u_1,u_2,\boldsymbol{x}), \quad u_1 = 1,\dots,k_1, u_2 = 1,\dots,k_2,$$

for the conditional probability of the response category y at occasion t, with $t = 1,\dots,T$ and $y = 0,\dots,c-1$, given the corresponding random effect U_1 and latent variable $U_2^{(t)}$.

The path diagram of the model based on the above assumptions is shown in Figure 6.1, where for simplicity we do not indicate the covariates.

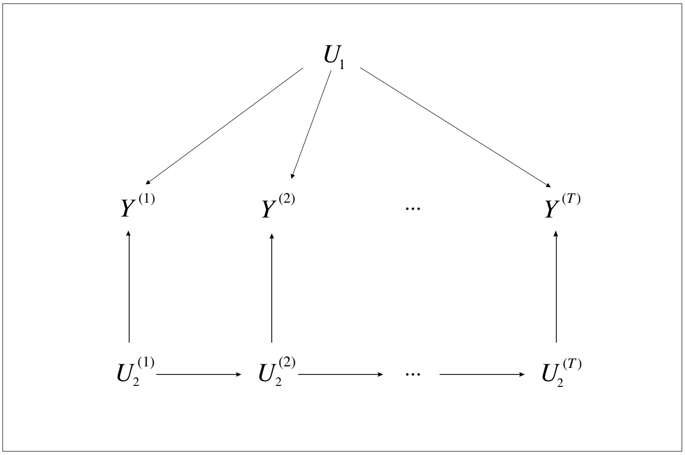

FIGURE 6.1
Path diagram for the LM model with the inclusion of random effects in the measurement model

This formulation, in which random effects (and covariates) are included in the measurement model, makes sense when we want to account for a further source of unobserved heterogeneity with respect to that represented by the latent Markov chain. The way in which the latent variable U_1 affects the distribution of $Y^{(t)}$, together with the covariates in $\boldsymbol{X}^{(t)}$ and $U_2^{(t)}$, may be formulated by employing the same parametrizations illustrated in Section 5.3. The following example, which extends Example 14, helps to clarify this case.

Example 21 — Logit model with time-invariant and time-varying random effects. *Consider the model for binary response vari-*

ables based on the following assumption:

$$\log \frac{\phi^{(t)}_{1|u_1 u_2 \boldsymbol{x}}}{\phi^{(t)}_{0|u_1 u_2 \boldsymbol{x}}} = \alpha_{1u_1} + \alpha_{2u_2} + \boldsymbol{x}'\boldsymbol{\psi},$$

where $t = 1, \dots, T$, $u_1 = 1, \dots, k_1$, and $u_2 = 1, \dots, k_2$. In the above parametrization, α_{1u_1} is an intercept corresponding to the support point u_1 for the latent variable U_1, α_{2u_2} is that corresponding to the latent state u_2 for the latent variable $U_2^{(t)}$, and $\boldsymbol{\psi}$ is a vector of regression parameters for the covariates. The resulting model then distinguishes between two forms of heterogeneity: time invariant (introduced by the parameters α_{1u_1}) and time varying (introduced by the parameters α_{2u_2}).

6.2.1.2 Random effects in the latent model

When the random effects are included in the latent model, the basic assumption is still that the variables in $\tilde{\boldsymbol{Y}}$ are conditionally independent given the latent process $\boldsymbol{U}_2$ and the covariates in $\tilde{\boldsymbol{X}}$, with every $Y^{(t)}$ depending only on $U_2^{(t)}$. However, the variables in $\boldsymbol{U}_2$ follow a first-order Markov chain only conditionally on U_1 and $\tilde{\boldsymbol{X}}$. See Figure 6.2 for an illustration via path diagram of the resulting LM model where, again, we omit indicating the covariates.

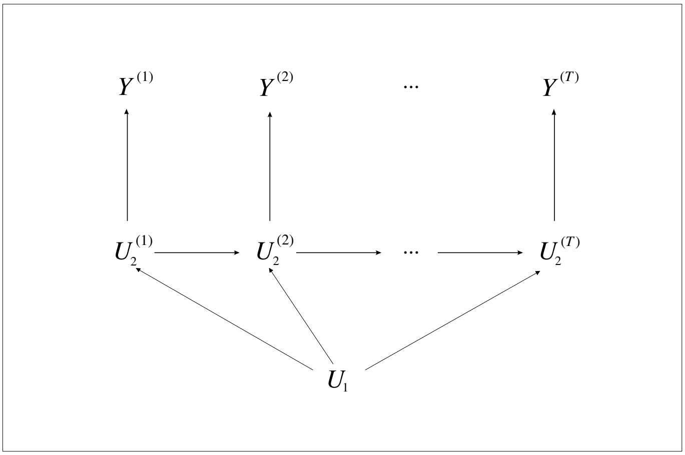

FIGURE 6.2
Path diagram for the LM model with the inclusion of random effects in the latent model

The extension at issue makes sense when we want to admit that the initial and the transition probabilities are different between subjects depending on observable and unobservable covariates. When U_1 is discrete, this approach may also be useful from a perspective of clustering. About the notation, we introduce the following symbols:

$$\begin{aligned}\pi_{u_2|u_1\boldsymbol{x}} &= f_{U_2^{(1)}|U_1,\boldsymbol{X}^{(1)}}(u_2|u_1,\boldsymbol{x}),\\ \pi^{(t)}_{u_2|u_1\bar{u}_2\boldsymbol{x}} &= f_{U_2^{(t)}|U_1,U_2^{(t-1)},\boldsymbol{X}^{(t)}}(u_2|u_1,\bar{u}_2,\boldsymbol{x}),\end{aligned}$$

with $t = 2,\dots,T$, $u_1 = 1,\dots,k_1$, and $\bar{u}_2, u_2 = 1,\dots,k_2$. These parameters correspond to the initial and the transition probabilities of the Markov chain associated with a subject with random effect u_1. Even in this case, suitable parametrizations, of the type illustrated in Section 5.4, have to be employed, as shown in the following example, which extends Example 16.

Example 22 — Tridiagonal transition matrix with global logit parametrization. *Assuming that the latent states are ordered, we allow the initial probabilities to depend on the U_1 and $\boldsymbol{X}^{(1)}$ through a global logit link function as follows:*

$$\log\frac{\pi_{u_2|u_1\boldsymbol{x}}+\cdots+\pi_{k_2|u_1\boldsymbol{x}}}{\pi_{1|u_1\boldsymbol{x}}+\cdots+\pi_{u_2-1|u_1\boldsymbol{x}}} = \gamma_{1u_1u_2}+\boldsymbol{x}'\boldsymbol{\gamma}_2,$$

for $u_1 = 1,\dots,k_1$ and $u_2 = 2,\dots,k_2$. Moreover, it is reasonable to assume that transition matrices are tridiagonal with transition probabilities parametrized as follows:

$$\log\frac{\pi^{(t)}_{u_2|u_1\bar{u}_2\boldsymbol{x}}}{\pi^{(t)}_{\bar{u}_2|u_1\bar{u}_2\boldsymbol{x}}} = \delta_{1u_1\bar{u}_2}+\delta_{2,(3+u_2-\bar{u}_2)/2}+\boldsymbol{x}'\boldsymbol{\delta}_3,$$

where $t = 2,\dots,T$, $u_1 = 1,\dots,k_1$, $\bar{u}_2 = 1,\dots,k_2$, and $u_2 = 2$ for $\bar{u}_2 = 1$, $u_2 = k_2 - 1$ for $\bar{u}_2 = k_2$, and $u_2 = \bar{u}_2 - 1, \bar{u}_2 + 1$ for $\bar{u}_2 = 2,\dots,k_2 - 1$.

The model resulting from the above parametrization allows us to spot k_1 clusters of subjects which are different in terms of initial and transition probabilities. The differences between clusters depend on the intercepts $\gamma_{1u_1u_2}$ and $\delta_{1u_1\bar{u}_2}$.

6.2.2 Manifest distribution

The manifest distribution of the response variables, that is, the conditional distribution of $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{X}}$, may be obtained by extending the rules given in Section 5.2 for LM models with covariates. In particular, the same rules as in (5.1), (5.2), and (5.3) may be applied to obtain the conditional distribution of $\boldsymbol{U}_2$ given U_1 and $\tilde{\boldsymbol{X}}$, that of $\tilde{\boldsymbol{Y}}$ given U_1, $\boldsymbol{U}_2$, and $\tilde{\boldsymbol{X}}$, and that of $\tilde{\boldsymbol{Y}}$ given U_1 and $\tilde{\boldsymbol{X}}$. More explicitly, we have

$$\begin{aligned}
f_{\boldsymbol{U}_2|U_1,\tilde{\boldsymbol{X}}}(\boldsymbol{u}_2|u_1,\tilde{\boldsymbol{x}}) &= \pi_{u_2^{(1)}|u_1\boldsymbol{x}^{(1)}}\prod_{t=2}^{T}\pi^{(t)}_{u_2^{(t)}|u_1u_2^{(t-1)}\boldsymbol{x}^{(t)}},\\
f_{\tilde{\boldsymbol{Y}}|U_1,\boldsymbol{U}_2,\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|u_1,\boldsymbol{u}_2,\tilde{\boldsymbol{x}}) &= \prod_{t=1}^{T}\phi^{(t)}_{y^{(t)}|u_1u_2^{(t)}\boldsymbol{x}^{(t)}},\\
f_{\tilde{\boldsymbol{Y}}|U_1,\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|u_1,\tilde{\boldsymbol{x}}) &= \sum_{\boldsymbol{u}_2}\pi_{u_2^{(1)}|u_1\boldsymbol{x}^{(1)}}\pi^{(2)}_{u_2^{(2)}|u_1u_2^{(1)}\boldsymbol{x}^{(2)}}\\
&\quad\times\cdots\pi^{(T)}_{u_2^{(T)}|u_1u_2^{(T-1)}\boldsymbol{x}^{(T)}}\\
&\quad\times\phi^{(1)}_{y^{(1)}|u_1u_2^{(1)}\boldsymbol{x}^{(1)}}\cdots\phi^{(T)}_{y^{(T)}|u_1u_2^{(T)}\boldsymbol{x}^{(T)}}.
\end{aligned}$$

In the above expressions, $\pi_{u_2^{(1)}|u_1\boldsymbol{x}^{(1)}} = \pi_{u_2^{(1)}}$ and $\pi^{(t)}_{u_2^{(t)}|u_1u_2^{(t-1)}\boldsymbol{x}^{(t)}} = \pi^{(t)}_{u_2^{(t)}|u_2^{(t-1)}}$ when covariates and random effects are only included in the measurement model. Similarly, $\phi^{(t)}_{y^{(t)}|u_1u_2^{(t)}\boldsymbol{x}^{(t)}} = \phi^{(t)}_{y^{(t)}|u_2^{(t)}}$ when covariates and random effects are only included in the latent model.

Once the distribution of $f_{\tilde{\boldsymbol{Y}}|U_1,\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|u_1,\tilde{\boldsymbol{x}})$ is computed for $u_1 = 1,\ldots,k_1$, the manifest distribution of $\tilde{\boldsymbol{Y}}$ is obtained by the sum

$$f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) = \sum_{u_1=1}^{k_1} f_{\tilde{\boldsymbol{Y}}|U_1,\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|u_1,\tilde{\boldsymbol{x}})\omega_{u_1}, \tag{6.1}$$

which depends on the mass probabilities for the distribution of the latent variable U_1. As usual, $f_{\tilde{\boldsymbol{Y}}|U_1,\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|u_1,\tilde{\boldsymbol{x}})$ is in practice computed by the forward recursion (Baum et al., 1970), already illustrated in the previous chapters.

Finally, note that if the random effects have a continuous distribution, an expression similar to expression (6.1) may be used to compute the manifest distribution. The main difference is that the sum over the support of U_1 is replaced by the sum over a suitable number of quadrature nodes, whereas the mass probabilities ω_{u_1} are replaced by suitable weights which possibly depend on the parameters of the random-effects distribution.

6.3 Maximum likelihood estimation

On the basis of data referred to an observed sample of size n, maximum likelihood estimation of a random-effects LM model, formulated as described in the previous section, is carried out by an extended version of the EM algorithm described in Section 5.6. We recall that the data consist of the vectors of covariates $\tilde{\boldsymbol{x}}_i$ and responses $\tilde{\boldsymbol{y}}_i$ observed for every sample unit i, with $i = 1, \ldots, n$. On the basis of these data, the log-likelihood may be expressed as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{x}}_i) = \sum_{\tilde{\boldsymbol{x}}} \sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} \log f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}),$$

where the sum $\sum_{\tilde{\boldsymbol{x}}} \sum_{\tilde{\boldsymbol{y}}}$ is extended to all pairs $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ of covariate and response vectors observed at least once, and as usual, $\boldsymbol{\theta}$ is the vector of all model parameters.

In order to describe the EM algorithm, we need to clarify the *complete data log-likelihood*, which has the following expression:

$$\begin{aligned}
\ell^*(\boldsymbol{\theta}) \quad = & \quad \sum_{u_1=1}^{k_1} \left[\sum_{t=1}^{T} \sum_{u_2=1}^{k_2} \sum_{\boldsymbol{x}} \sum_{y=0}^{c-1} a_{u_1 u_2 \boldsymbol{x} y}^{(t)} \log \phi_{y|u_1 u_2 \boldsymbol{x}}^{(t)} \right. \\
& + \quad \sum_{u_2=1}^{k_2} \sum_{\boldsymbol{x}} b_{u_1 u_2 \boldsymbol{x}}^{(1)} \log \pi_{u_2|u_1 \boldsymbol{x}} \\
& + \quad \left. \sum_{t=2}^{T} \sum_{\bar{u}_2=1}^{k_2} \sum_{u_2=1}^{k_2} \sum_{\boldsymbol{x}} b_{u_1 \bar{u}_2 u_2 \boldsymbol{x}}^{(t)} \log \pi_{u_2|u_1 \bar{u}_2 \boldsymbol{x}}^{(t)} \right] \\
& + \quad \sum_{u_1=1}^{k_1} c_{u_1} \log \omega_{u_1}, \qquad (6.2)
\end{aligned}$$

where, for every u_1, the frequencies $a_{u_1 u_2 \boldsymbol{x} y}^{(t)}$, $b_{u_1 u_2 \boldsymbol{x}}^{(t)}$, and $b_{u_1 \bar{u}_2 u_2 \boldsymbol{x}}^{(t)}$ are defined as in Section 5.6.1. In particular, $a_{u_1 u_2 \boldsymbol{x} y}^{(t)}$ is the number of sample units that are in latent class u_1 (namely for which $U_1 = u_1$) and, at occasion t, are in latent state u_2, have covariates $\boldsymbol{x}$, and provide response y. With reference to the same time occasion t, latent class u_1, and covariate configuration $\boldsymbol{x}$, $b_{u_1 u_2 \boldsymbol{x}}^{(t)}$ is the number of sample units in latent state u_2 and $b_{u_1 \bar{u}_2 u_2 \boldsymbol{x}}^{(t)}$ is the number of transitions from latent state $\bar{u}_2$ to u_2. Finally, c_{u_1} is the overall number of sample units that are in latent class u_1.

The EM algorithm alternates the following two steps until convergence:

- **E-step**: compute the expected value of the above frequencies, given the observed data and the current value of the parameters, so as to obtain the expected value of $\ell^*(\boldsymbol{\theta})$. The expected values of these frequencies may be expressed as follows:

$$\begin{aligned}
&\hat{a}^{(t)}_{u_1u_2\boldsymbol{x}y}\\
&\quad=\sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} f_{U_1,U_2^{(t)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u_1,u_2|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) I(\boldsymbol{x}^{(t)}=\boldsymbol{x},y^{(t)}=y),\\
&\hat{b}^{(t)}_{u_1u_2\boldsymbol{x}}\\
&\quad=\sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} f_{U_1,U_2^{(t)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u_1,u_2|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) I(\boldsymbol{x}^{(t)}=\boldsymbol{x}),\\
&\hat{b}^{(t)}_{u_1\bar{u}_2u_2\boldsymbol{x}}\\
&\quad=\sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} f_{U_1,U_2^{(t-1)},U_2^{(t)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u_1,\bar{u}_2,u_2|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) I(\boldsymbol{x}^{(t)}=\boldsymbol{x}),\\
&\hat{c}_{u_1}\\
&\quad=\sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} f_{U_1|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u_1|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}),
\end{aligned}$$

  with $t=1,\ldots,T$, $u_1=1,\ldots,k_1$, $\bar{u}_2,u_2=1,\ldots,k_2$, and $y=0,\ldots,c-1$. These expected frequencies involve the posterior distribution of the time-invariant and time-varying latent variables. In this regard, consider that the posterior distribution of U_1, used to compute $\hat{c}_{u_1}$, is simply given by

$$f_{U_1|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u_1|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})=\frac{f_{\tilde{\boldsymbol{Y}}|U_1,\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|u_1,\tilde{\boldsymbol{x}})\omega_{u_1}}{f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}})},$$

  where in the denominator we have the manifest probability defined in (6.1). Moreover, about the posterior distribution involved in the expressions for $\hat{a}^{(t)}_{u_1u_2\boldsymbol{x}y}$ and $\hat{b}^{(t)}_{u_1u_2\boldsymbol{x}}$, consider that

$$\begin{aligned}
&f_{U_1,U_2^{(t)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u_1,u_2|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\\
&\quad=f_{U_1|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u_1|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) f_{U_2^{(t)}|U_1,\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u_2|u_1,\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}),
\end{aligned}$$

  where $f_{U_2^{(t)}|U_1,\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u_2|u_1,\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})$ is obtained by the usual recursions. In a similar way we may express the posterior distribution required to compute the expected frequencies $\hat{b}^{(t)}_{u_1\bar{u}_2u_2\boldsymbol{x}}$.

- **M-step**: maximize the complete data log-likelihood, as expressed in (6.2), with each frequency replaced by the corresponding expected value. The maximization is performed separately for each block of parameters, as shown in the following:

- *Conditional response probabilities*: if these probabilities do not depend either on individual covariates or on random effects, they are updated on the basis of the rules given in Sections 3.5.1 and 4.4.1. For instance, if the only constraint is that of time homogeneity of these probabilities, we compute

  $$\phi_{y|u_2} = \frac{\sum_{t=1}^{T} \sum_{u_1=1}^{k_1} \sum_{\boldsymbol{x}} \hat{a}_{u_1 u_2 \boldsymbol{x} y}^{(t)}}{\sum_{t=1}^{T} \sum_{u_1=1}^{k_1} \sum_{\boldsymbol{x}} \hat{b}_{u_1 u_2 \boldsymbol{x}}^{(t)}},$$

  for $u_2 = 1, \dots, k_2$ and $y = 0, \dots, c-1$. If the conditional response probabilities are modeled on the basis of the covariates and random effects, then an iterative algorithm of the type described in Section 5.6.1 is necessary to update the corresponding parameters.

- *Initial probabilities*: if these probabilities are not assumed to depend either on covariates or on random effects, but are unconstrained, then they are updated by a standard rule; in practice, we compute

  $$\pi_{u_2} = \frac{\sum_{u_1=1}^{k_1} \sum_{\boldsymbol{x}} \hat{b}_{u_1 u_2 \boldsymbol{x}}^{(1)}}{n}, \quad u_2 = 1, \dots, k_2.$$

  Otherwise, these probabilities are updated as described in Section 5.6.1 when they depend on covariates and random effects.

- *Transition probabilities*: if the transition probabilities do not depend either on covariates or on random effects, then they are updated through the rules described in Sections 3.5.1 and 4.4.1, according to the way in which they are modeled. For instance, if they are time heterogenous, we have to compute

  $$\pi_{u_2|\bar{u}_2}^{(t)} = \frac{\sum_{u_1=1}^{k_1} \sum_{\boldsymbol{x}} \hat{b}_{u_1 \bar{u}_2 u_2 \boldsymbol{x}}^{(t)}}{\sum_{u_1=1}^{k_1} \sum_{\boldsymbol{x}} \hat{b}_{u_1 \bar{u}_2 \boldsymbol{x}}^{(t-1)}},$$

  for $t = 2, \dots, T$ and $\bar{u}_2, u_2 = 1, \dots, k_2$. Finally, when the individual covariates and random effects are assumed to affect the transition probabilities, an iterative algorithm of the type described in Section 5.6.1 is necessary.

- *Mass probabilities*: the parameters of the distribution of the latent variable U_1 are simply updated as

  $$\omega_{u_1} = \frac{\hat{c}_{u_1}}{n}, \quad u_1 = 1, \dots, k_1.$$

Essentially, the same EM algorithm as above may be used in the presence of random effects having a continuous distribution. However, the algorithm is slightly more complex due to the substitution in (6.2) of the sums over the support of U_1 by a quadrature over a large number of nodes; see also the comment at the end of Section 6.2.2.

6.4 Multilevel formulation

Multilevel LM models are useful when sample units are collected in a certain number H of clusters and we want to take into account a cluster effect on the responses provided by these units. In the following, we illustrate the LM multilevel version proposed by Bartolucci et al. (2011) which also includes individual and cluster-level covariates. A similar model was proposed by Yu (2008), whereas a version based on random effects having a continuous distribution was developed by Asparouhov and Muthén (2008).

6.4.1 Model assumptions

In order to illustrate the multilevel extension of the LM model, we have to change the notation used previously making explicit reference to the subject and the cluster to which certain variables are referred. In particular, every subject in the sample is identified by the pair of indices hi, with $h = 1, \dots, H$ and $i = 1, \dots, n_h$, where n_h is the dimension of cluster h. Accordingly, we denote the response of this subject at occasion t by $Y_{hi}^{(t)}$, the vector of the overall responses provided by this subject by $\tilde{\boldsymbol{Y}}_{hi}$, and the collection of all responses provided by the subjects in cluster h by $\tilde{\boldsymbol{Y}}_h$. Moreover, for every subject hi we assume the existence of a latent process $\boldsymbol{U}_{2hi} = (U_{2hi}^{(1)}, \dots, U_{2hi}^{(T)})$ that follows a Markov chain with state space $1, \dots, k_2$, whereas, for each cluster h, we introduce the latent variable U_{1h}. Finally, note that in this context we have covariates at both cluster and individual levels. The covariates of the first type are collected in the vectors $\boldsymbol{X}_{1h}$, whereas those at the individual level are collected in the vectors $\boldsymbol{X}_{2hi}^{(1)}, \dots, \boldsymbol{X}_{2hi}^{(T)}$, for $h = 1, \dots, H$ and $i = 1, \dots, n_h$. All covariate vectors at the individual level are collected in $\tilde{\boldsymbol{X}}_{2hi}$, which has a structure similar to the vector $\tilde{\boldsymbol{X}}$ previously used, whereas all covariates for the units in the same cluster h are included in the vector $\tilde{\boldsymbol{X}}_{2h}$.

As usual, the effect of the clusters, represented by the latent variables U_{1h}, may be included in the measurement model or in the latent model. Moreover, these latent variables may be assumed to have either a continuous or a discrete distribution. However, in the following we explicitly illustrate only the case of discrete cluster-level latent variables, with k_1 support points, entering in the latent model. The illustration closely follows that provided in Bartolucci et al. (2011), to which we refer the reader for details.

When the cluster-level latent variables U_{1h} are included in the latent model, the assumption of local independence is retained and formulated by requiring that the response variables in $\tilde{\boldsymbol{Y}}_{hi}$ are conditionally independent, given the latent process in $\boldsymbol{U}_{2hi}$ and the covariates in $\tilde{\boldsymbol{X}}_{2hi}$. Moreover, every response variable $Y_{hi}^{(t)}$ depends only on $U_{2hi}^{(t)}$, and the latent variables in $\boldsymbol{U}_{2hi}$ are assumed to follow a Markov chain conditionally on U_{1h} and $\tilde{\boldsymbol{X}}_{2hi}$. Finally, it is assumed that the cluster-level covariates in $\boldsymbol{X}_{1h}$ directly affect the distribution of the latent variable U_{1h}.

The model based on the above assumptions is represented by the path diagram in Figure 6.3, in which we omit indicating the cluster- and individual-level covariates for simplicity.

The above assumptions imply that, for every h, the response vectors $\tilde{\boldsymbol{Y}}_{h1}, \ldots, \tilde{\boldsymbol{Y}}_{hn_h}$ are not independent, but they are conditionally independent given the latent variable U_{1h}. Marginal independence holds between the vectors of response variables $\tilde{\boldsymbol{Y}}_1, \ldots, \tilde{\boldsymbol{Y}}_H$ associated with the different clusters. Taking into account these assumptions, we formulate the basic notation as follows. The conditional response probabilities are denoted by

$$\phi_{y|u_2} = f_{Y_{hi}^{(t)}|U_{2hi}^{(t)}}(y|u_2), \quad t = 1, \ldots, T, \; u_2 = 1, \ldots, k_2, \; y = 0, \ldots, c-1, \tag{6.3}$$

for all sample units hi, when these probabilities are assumed to be time homogeneous. Moreover, the initial probabilities of the latent process are still denoted as follows:

$$\pi_{u_2|u_1\boldsymbol{x}_2} = f_{U_{2hi}^{(1)}|U_{1h},\boldsymbol{X}_{2hi}^{(t)}}(u_2|u_1, \boldsymbol{x}_2), \quad u_1 = 1, \ldots, k_1, \;\; u_2 = 1, \ldots, k_2, \tag{6.4}$$

for every cluster h and unit hi in the cluster. Similarly, the transition probabilities of the latent process are defined as

$$\pi_{u_2|u_1\bar{u}_2\boldsymbol{x}_2}^{(t)} = f_{U_{2h}^{(t)}|U_{1h},U_{2hi}^{(t-1)},\boldsymbol{X}_{2hi}^{(t)}}(u_2|u_1, \bar{u}_2, \boldsymbol{x}_2), \tag{6.5}$$

where $t = 2, \ldots, T$, $u_1 = 1, \ldots, k_1$, and $\bar{u}_2, u_2 = 1, \ldots, k_2$. Finally, with reference to the distribution of the cluster-level latent variables, we introduce the probabilities

$$\omega_{u_1|\boldsymbol{x}_1} = f_{U_{1h}|\boldsymbol{X}_{1h}}(u_1|\boldsymbol{x}_1), \quad h = 1, \ldots, H, \; u_1 = 1, \ldots, k_1, \tag{6.6}$$

which are conditional on the cluster-level covariates. Note that these covariates are not time varying.

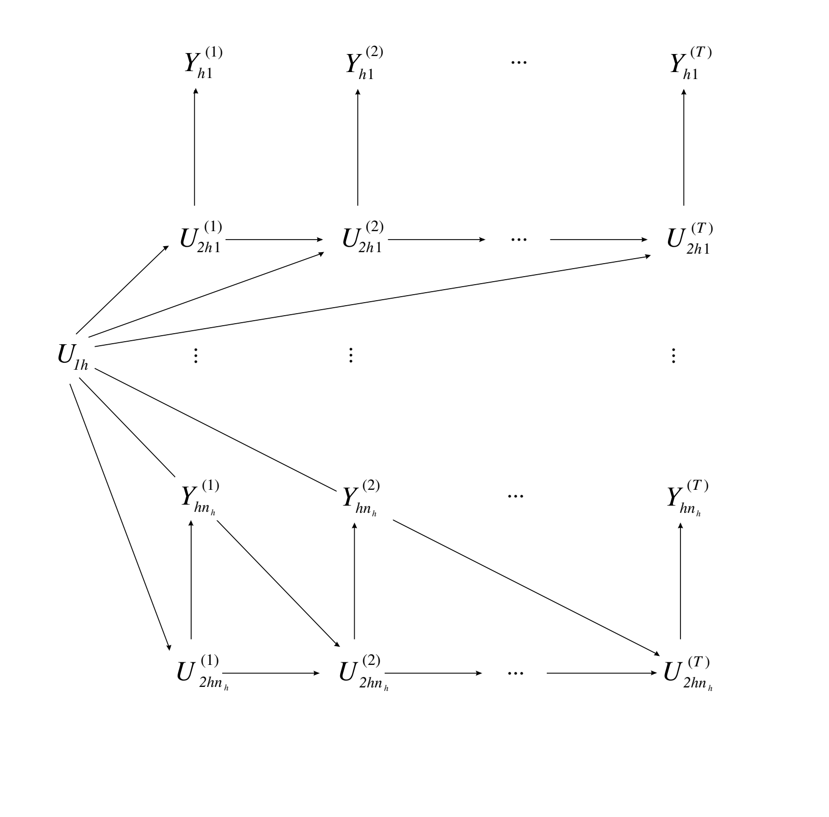

FIGURE 6.3
Path diagram for the multilevel LM model with random effects in the latent model

In practice, the probabilities in (6.4), (6.5), and (6.6) depend on the corresponding covariates by suitable parametrizations of the type described in Section 5.4. This is clarified in the following example.

Example 23 — Multilevel LM model with individual and cluster level covariates. *Suppose that the latent states have an ordinal nature because, for instance, we assume that the conditional response probabilities follow a Rasch model; see Example 10. Then, we have*

$$\log \frac{\phi_{1|u_2}^{(t)}}{\phi_{0|u_2}^{(t)}} = \alpha_{u_2} - \psi^{(t)}, \qquad t = 1, \ldots, T,\ u_2 = 1, \ldots, k_2, \qquad (6.7)$$

where α_{u_2} is interpreted as the ability level of the subjects in latent state u_2, whereas $\psi^{(t)}$ is interpreted as the difficulty level of item t. Note that, differently from assumption (6.3), the conditional response probabilities are not time homogenous. Then, initial probabilities of every individual Markov chain may be parametrized by global logits as follows:

$$\log \frac{\pi_{u_2|u_1\boldsymbol{x}_2} + \cdots + \pi_{k_2|u_1\boldsymbol{x}_2}}{\pi_{1|u_1\boldsymbol{x}_2} + \cdots + \pi_{u_2-1|u_1\boldsymbol{x}_2}} = \gamma_{1u_1} + \gamma_{2u_2} + \boldsymbol{x}_2'\boldsymbol{\gamma}_3, \tag{6.8}$$

for $u_1 = 1, \ldots, k_1$, and $u_2 = 2, \ldots, k_2$. A similar parametrization may be adopted for the transition probabilities, by requiring that

$$\log \frac{\pi^{(t)}_{u_2|u_1\bar{u}_2\boldsymbol{x}_2} + \cdots + \pi^{(t)}_{k_2|u_1\bar{u}_2\boldsymbol{x}_2}}{\pi^{(t)}_{1|u_1\bar{u}_2\boldsymbol{x}_2} + \cdots + \pi^{(t)}_{u_2-1|u_1\bar{u}_2\boldsymbol{x}_2}} = \delta^{(t)}_{1u_1} + \delta^{(t)}_{2\bar{u}_2u_2} + \boldsymbol{x}_2'\boldsymbol{\delta}^{(t)}_3, \tag{6.9}$$

for $t = 2, \ldots, T$, $u_1 = 1, \ldots, k_1$, $\bar{u}_2 = 1, \ldots, k_2$, and $u_2 = 2, \ldots, k_2$.

Note that the parameters γ_{1u_1} and $\delta^{(t)}_{1u_1}$ have a special role, since they measure the effect of the cluster on the initial and transition probabilities. Suppose, for instance, that the response variables measure the level of a certain ability of students collected in classes. Then, the initial ability of the students in a class of typology u_1 increases with γ_{1u_1}. Similarly, with reference to the time occasion t, we have larger probabilities of the transition from a level of ability to the higher level as the parameter $\delta^{(t)}_{1u1}$ increases.

Finally, the effect of the clusters may depend on the corresponding covariates, such as the type of school or teaching method, which are included into $\boldsymbol{X}_{1h}$ by suitable dummy variables. Then, we may use a multinomial logit parametrization of the mass probabilities $\omega_{u_1|\boldsymbol{x}_1}$. In this case, we assume:

$$\log \frac{\omega_{u_1|\boldsymbol{x}_1}}{\omega_{1|\boldsymbol{x}_1}} = \tau_{1u_1} + \boldsymbol{x}_1'\boldsymbol{\tau}_{2u_1}, \quad u_1 = 2, \ldots, k_1. \tag{6.10}$$

Therefore, once we have described the different typologies of clusters on the basis of the parameters γ_{1u_1} and $\delta^{(t)}_{1u_1}$, we can assess how the cluster-level covariates affect the probability of being of a certain typology.

6.4.2 Manifest distribution and maximum likelihood estimation

For a given subject hi in cluster h, the conditional distribution of the response variables in $\tilde{\boldsymbol{Y}}_{hi}$, given the corresponding covariates in $\tilde{\boldsymbol{X}}_{2hi}$ and U_1, may be obtained as clarified in Section 6.2.2 about the random-effects LM models. This distribution is denoted by $f_{\tilde{\boldsymbol{Y}}_{hi}|U_{1h},\tilde{\boldsymbol{X}}_{2hi}}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{x}}_2)$. Then, by an expression similar to (6.1) we may obtain the manifest distribution of $\tilde{\boldsymbol{Y}}_{hi}$ given the covariates in $\boldsymbol{X}_{1h}$ and $\tilde{\boldsymbol{X}}_{2hi}$, that is,

$$f_{\tilde{\boldsymbol{Y}}_{hi}|\boldsymbol{X}_{1h},\tilde{\boldsymbol{X}}_{2hi}}(\tilde{\boldsymbol{y}}|\boldsymbol{x}_1,\tilde{\boldsymbol{x}}_2) = \sum_{u_1=1}^{k_1} f_{\tilde{\boldsymbol{Y}}_{hi}|U_{1h},\tilde{\boldsymbol{X}}_{2hi}}(\tilde{\boldsymbol{y}}|u_1,\tilde{\boldsymbol{x}}_2)\omega_{u_1|\boldsymbol{x}_1}.$$

As clarified above, sample units in the same cluster are independent only conditionally on the corresponding cluster-level latent variable and covariates. Then, the manifest distribution for the responses provided by all the units in the same cluster h is obtained as

$$f_{\tilde{\boldsymbol{Y}}_{h}|\boldsymbol{X}_{1h},\tilde{\boldsymbol{X}}_{2h}}(\tilde{\boldsymbol{y}}|\boldsymbol{x}_1,\tilde{\boldsymbol{x}}_2) = \sum_{u_1=1}^{k_1} \omega_{u_1|\boldsymbol{x}_1}\prod_{i=1}^{n_h} f_{\tilde{\boldsymbol{Y}}_{hi}|U_{1h},\tilde{\boldsymbol{X}}_{2hi}}(\tilde{\boldsymbol{y}}_i|u_1,\tilde{\boldsymbol{x}}_{2i}),$$

where $\tilde{\boldsymbol{y}}_i$ is the subvector of $\tilde{\boldsymbol{y}}$ containing the responses referred to unit i, and similarly, $\tilde{\boldsymbol{x}}_{2i}$ is a subvector of $\tilde{\boldsymbol{x}}_2$, with $i = 1, \dots, n_h$.

The distribution above is the basis for the model log-likelihood, which is used for parameter estimation. Note that, in this context, the sample data correspond to the vectors of cluster-level covariates $\boldsymbol{x}_{1h}$, $h = 1, \dots, H$, and to the vectors of individual-level covariates and responses, that is $\tilde{\boldsymbol{x}}_{2hi}$ and $\tilde{\boldsymbol{y}}_{hi}$, for $h = 1, \dots, H$ and $i = 1, \dots, n_h$, which are collected in $\tilde{\boldsymbol{x}}_{2h}$ and $\tilde{\boldsymbol{y}}_h$, respectively. The corresponding log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{h=1}^{H} \log f_{\tilde{\boldsymbol{Y}}_{h}|\boldsymbol{X}_{1h},\tilde{\boldsymbol{X}}_{2h}}(\tilde{\boldsymbol{y}}_h|\boldsymbol{x}_{1h},\tilde{\boldsymbol{x}}_{2h}),$$

which is based on the independence between the responses and the corresponding latent variables referred to different clusters.

In order to maximize the above log-likelihood and then obtain the maximum likelihood estimate of the model parameters collected in $\boldsymbol{\theta}$, we can use an EM algorithm similar to the one illustrated in Section 6.3. However, we prefer to avoid an explicit description of the algorithm and refer the reader to Bartolucci et al. (2011) for details.

6.5 Application to the student math achievement dataset

In order to illustrate the random-effects extension of LM models, and in particular the version for multilevel data outlined in Section 6.4, we summarize the results of the application of this model to the math achievement dataset introduced in Section 1.4.4. For a detailed description of this application we refer the reader to Bartolucci et al. (2011). We recall that these data were collected on a large number of middle-school students, grouped in classes belonging to different schools, at the end of each year of schooling from Grade 6 through Grade 8, so that $T = 3$.

The model applied for this analysis allows us to explain the dynamics of achievement, the heterogeneity between subjects, and measurement errors in the test administration, taking into account the different nature of the schools. We mainly distinguish between public and nonpublic schools. In particular, the model is a multivariate extension of that formulated in Example 23. Since several items are administered at the end of each year of schooling, we have more response variables for each occasion. Then, we have a series of difficulty parameters, one for every distinct item used in the questionnaire, instead of a single difficulty level for each time occasion as in (6.7). However, the model parameters may be interpreted as clarified in that example.

As for standard LM models, a crucial point in the application of the multilevel LM model at issue is the choice of the number of latent states, here denoted by k_2. In this case we also have to select the number of support points for every cluster-level latent variable (k_1). For this aim, Bartolucci et al. (2011) adopted the Bayesian information criterion, on the basis of which they selected the model with $k_1 = 4$ support points for the latent variable at cluster level, which correspond to four *typologies of class of students*, and $k_2 = 6$ states for the latent Markov chain at individual level.

In commenting the estimates of the parameters of the model selected as above, we first consider the ability parameters for every latent state (α_{u_2}). These estimates are reported in Table 6.1, which shows that the latent states correspond to increasing levels of math ability of the students (the first level is constrained to 0 to ensure model identifiability).

Then, Table 6.2 displays the estimates of the parameters of most interest for the evaluation of the school effect. These parameters affect the initial and transition probabilities of the individual Markov chain for

TABLE 6.1
Estimates of the math ability levels α_{u_2} under the selected model

u_2	$\hat{\alpha}_{u_2}$
1	0.000
2	0.866
3	1.800
4	2.698
5	3.623
6	4.825

TABLE 6.2
Estimates of the parameters γ_{1u_1} and $\delta_{1u_1}^{(t)}$ which measure the cluster effect on the initial and transition probabilities

		$\hat{\delta}_{1u_1}^{(t)}$	
u_1	$\hat{\gamma}_{1u_1}$	$t = 2$	$t = 3$
1	0.000	0.000	0.000
2	−0.261	−2.774	3.893
3	1.140	−0.317	1.673
4	0.013	4.515	−0.434

the math ability through (6.8) and (6.9). In particular, there are four intercepts for each type of probability, which correspond to the effect of the different typologies of class of students. Related parameters are regression coefficients for the individual covariates corresponding to the educational level of the student's parents. These estimates are given in Bartolucci et al. (2011).

On the basis of the results in Table 6.2, we conclude that classes of typology 2 have the lowest effect on the math ability at the first year, whereas classes of typology 3 have the highest effect. Concerning the parameters affecting the transition probabilities of the latent Markov chain, from Grade 6 to Grade 7 ($t = 2$), we conclude that classes of typology 2 contribute the least to the improvement of the math ability, whereas classes of typology 4 contribute the most. Opposite effects are observed for the transition from Grade 7 to Grade 8 ($t = 3$). In any case, classes of typology 1 and 3 always have an intermediate effect in terms of improvement in the math ability.

In order to better interpret the above estimates, in Table 6.3 we show the corresponding average ability level for each grade and typology of class and its increase with respect to the previous grade. These results show that classes of typology 2 are the least helpful in increasing math ability in the first year of middle school, as compared to the classes of the other typologies. The same classes contribute the least also from Grade 6 to Grade 7 but they contribute the most to students' math abilities from Grade 7 to 8.

Finally, the estimates of the parameters of the logistic model for the distribution of each cluster-level latent variable, which is based on parametrization (6.10), are reported in Table 6.4. In particular, the parameters τ_{1u_1} and those in the vector $\boldsymbol{\tau}_{2u_1}$ are referred to the logit comparing the probability of typology u_1 with the probability of typology 1, where $u_1 = 2, 3, 4$. We recall that τ_{1u_1} is an intercept, whereas $\boldsymbol{\tau}_{2u_1}$ contains the regression coefficients for the dummy variable *type of school*,

TABLE 6.3
Estimated average ability level of a reference student and its increase with respect to the previous grade according to the typology of class

Typology (u_1)	Grade	Average ability	Increase
1	6	1.046	–
	7	2.275	1.229
	8	2.724	0.449
2	6	0.944	–
	7	1.576	0.632
	8	2.973	1.397
3	6	1.521	–
	7	2.572	1.051
	8	3.358	0.786
4	6	1.051	–
	7	3.421	2.370
	8	3.555	0.134

TABLE 6.4
Estimates of the parameters τ_{1u_1} and $\boldsymbol{\tau}_{2u_1}$ in the model for the distribution of each cluster-level latent variable

		$\hat{\boldsymbol{\tau}}_{2u1}$		
u_1	$\hat{\tau}_{1u_1}$	Type	Stud./Teach.	Years
2	-42.030	28.815	1.263	-1.283^{**}
3	-31.359	28.389	0.336^{**}	-2.851^{**}
4	2.812	-0.039^{**}	-0.620	2.754

Note: ** significant at the 5% level

for the dummy variable *students/teachers ratio* (less than or equal to 8 or greater than 8), and for the dummy variable *years of activity of the school* (greater than or equal to 17.5 years or less than 17.5 years).

On the basis of the parameter estimates for the type of school, we conclude that public schools have a much higher probability of being of typology 2 or 3, with respect to typology 1 or 4. Similarly, the second covariate positively affects the probability that the school is of typology 2 or 3. On the other hand, for schools with fewer years of activity, the probability of typology 2 and 3 tends to be smaller and the probability of typology 4 tends to be higher. In order to have a clear description of the effect of these covariates, Table 6.5 reports the percentage frequency of the schools assigned to each typology according to the corresponding posterior distribution, which is based on the cluster-level covariates.

TABLE 6.5
Percentage frequencies of schools assigned to each typology according to the type (public/nonpublic), students/teachers ratio, and years since the school opened

		u_1			
		1	2	3	4
Type	nonpublic	78.57	0.00	0.00	21.43
	public	31.75	25.40	36.51	6.35
Stud./Teach.	≤ 8	66.67	4.76	4.76	23.81
	> 8	30.36	26.79	39.29	3.57
Years	$\geq$ 17.5 years	42.11	21.05	29.83	7.01
	$<$ 17.5 years	35.00	20.00	30.00	15.00

We observe that typology 4 mainly characterizes nonpublic schools with fewer than 18 years since the school opened and with a ratio between students and teachers smaller than 8. As this ratio increases, there is more chance that the class is of typology 1 rather than of typology 2. As the value of the year of activity of the school increases there is more chance that the class is of typology 1 rather than of typology 2.

7

Advanced topics about latent Markov modeling

7.1 Introduction

In this chapter, we deal with miscellaneous advanced topics about latent Markov (LM) models. First of all, we introduce models for continuous responses, formulating assumptions on either the conditional mean or a conditional quantile of the distribution of the response variable, given the corresponding covariates and latent variables. Then, we discuss how to deal with missing responses and we detail better certain computational issues: the Newton-Raphson (NR) algorithm for computing maximum likelihood estimates of the parameters and the parametric bootstrap to obtain the corresponding standard errors.

A further problem related to estimation, that we discuss in detail, is decoding, that is, prediction of the hidden state at each occasion on the basis of the observed data; this is discussed together with the problem of forecasting. A further crucial issue concerns the choice of the number of latent states. In this regard, we discuss and compare common information criteria by a simulation study.

7.2 Dealing with continuous response variables

A remarkable feature of LM models, and also of hidden Markov (HM) models, is that practically any type of response variable may be handled with minor adjustments to the general framework. In fact, it is only necessary to reformulate the measurement model by appropriate assumptions on the conditional distribution of the response variables, given the corresponding latent variables and possible covariates.

We now discuss the case in which the response variables in $\tilde{\boldsymbol{Y}}$ are continuous. In these cases it is common to also have individual covariates.

We then introduce the notation

$$f^{(t)}(y|u,\boldsymbol{x}) = f_{Y^{(t)}|U^{(t)},\boldsymbol{X}^{(t)}}(y|u,\boldsymbol{x})$$

for the density function of the response variable $Y^{(t)}$ given $U^{(t)}$ and $\boldsymbol{X}^{(t)}$; this function will depend on specific parameters. For simplicity, we restrict the illustration to univariate models with covariates affecting only the measurement model, but we show two different ways of formulating the conditional distribution of the response variables given the latent variables and the covariates. The first formulation is based on a standard linear regression approach, whereas the second is based on a quantile regression (QR) approach.

Obviously, even for the case of continuous response variables, the same formula as in (3.2) may be used to express the probability mass function of the distribution of $\boldsymbol{U}$, denoted by $f_{\boldsymbol{U}}(\boldsymbol{u})$, on the basis of the parameters π_u and $\pi^{(t)}_{u|\bar{u}}$. On the other hand, an expression similar to (5.2) may be used to compute the density function $f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U}}(\tilde{\boldsymbol{y}}|\boldsymbol{u})$. Finally, adapting expression (5.3) we obtain the density function of the manifest distribution of all response variables in $\tilde{\boldsymbol{Y}}$, which is equal to

$$\begin{aligned} f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) &= \sum_{\boldsymbol{u}} \pi_{u^{(1)}} \pi^{(2)}_{u^{(2)}|u^{(1)}} \cdots \pi^{(T)}_{u^{(T)}|u^{(T-1)}} \\ &\quad \times f^{(1)}(y^{(1)}|u^{(1)},\boldsymbol{x}^{(1)}) \cdots f^{(T)}(y^{(T)}|u^{(T)},\boldsymbol{x}^{(T)}). \end{aligned}$$

In practice, this density function is computed by the usual forward recursion, which may be efficiently implemented by the matrix notation, as clarified in Chapter 3. In particular, we have the same expression as in (3.5) for the density function at issue, whereas in the recursions in (3.6) and (3.7) we have to replace $\phi_{y^{(t)}|u}$ with $f^{(t)}(y|u,\boldsymbol{x})$.

7.2.1 Linear regression

In this case we assume that, conditionally on the latent variable and the covariates, each response variable has a normal distribution with a specific mean, denoted by $\eta^{(t)}_{u\boldsymbol{x}}$, and common variance σ^2. More precisely, we assume that

$$f^{(t)}(y|u,\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y-\eta^{(t)}_{u\boldsymbol{x}})^2}{\sigma^2}\right],$$

where

$$\eta^{(t)}_{u\boldsymbol{x}} = (\boldsymbol{w}^{(t)}_{u\boldsymbol{x}})'\boldsymbol{\beta}, \tag{7.1}$$

with $\boldsymbol{w}^{(t)}_{u\boldsymbol{x}}$ being a vector depending on the covariates in $\boldsymbol{x}$ and $\boldsymbol{\beta}$ being a specific vector of regression parameters. This formulation closely recalls that for categorical data which is based on (5.5), but here we do not need to introduce a link function.

The example below helps to clarify how an LM model may be formulated following this approach.

Example 24 — Switching regression model. *A natural formulation for the conditional mean of each response variable is*

$$\eta_{u\boldsymbol{x}}^{(t)} = \alpha_u + \boldsymbol{x}'\boldsymbol{\psi},$$

where α_u is the intercept for each latent state u. This assumption may be expressed as in (7.1), with

$$\boldsymbol{w}_{u\boldsymbol{x}}^{(t)} = \begin{pmatrix}\boldsymbol{d}_{uk}' & \boldsymbol{x}'\end{pmatrix}', \quad \boldsymbol{\beta} = \begin{pmatrix}(\alpha_1, \ldots, \alpha_k) & \boldsymbol{\psi}'\end{pmatrix}'.$$

A switching regression model (e.g., Hamilton, 1989) for longitudinal data results, in which the intercept is time varying.

It is also interesting to note that, in the absence of individual covariates, an extension of a finite-mixture model (McLachlan and Peel, 2000) results, in which sample units may move between the clusters corresponding to the different components. In this case, the measurement model only depends on the parameters α_u, $u = 1, \ldots, k$, with α_u equal to the conditional expected value of $Y^{(t)}$ given $U^{(t)} = u$.

7.2.2 Quantile regression

A somewhat different situation arises when we want to model a quantile of the continuous response. QR (Koenker and Bassett, 1978) is an approach to regression in which the effect of the covariates is modeled on a conditional quantile of the response variable, rather than on the conditional mean. This approach can be used to describe the entire distribution of a response, rather than just its expected value. A clear advantage of the QR, with respect to classical regression, is that it is robust with respect to outliers. Even when the center of the distribution is of main interest, one can model the 50th percentile, that is, perform a median regression. QR has been applied in many fields; see, for instance, Yu et al. (2003), Koenker (2005), and the references therein. There are many extensions of this type of regression for longitudinal data (Jung, 1996; Lipsitz et al., 1997; Koenker, 2004; Geraci and Bottai, 2007; Liu and Bottai, 2009), but most of them are confined to time-constant random effects. A time-varying approach to QR can be implemented within the LM framework. This was proposed by Farcomeni (2012), and it is illustrated in details in the following.

For a fixed probability level $\tau \in (0, 1)$, the approach consists of modeling the corresponding quantile of the conditional distribution of $Y^{(t)}$

given $U^{(t)}$ and $\boldsymbol{X}^{(t)}$. In this regard, a convenient parametric assumption is that this distribution is of asymmetric Laplace type, that is,

$$f^{(t)}(y|u,\boldsymbol{x}) = \frac{\tau(1-\tau)}{\sigma}\exp\left[-\rho_\tau\left(\frac{y-\eta_{u\boldsymbol{x}}^{(t)}}{\sigma}\right)\right],$$

where $\eta_{u\boldsymbol{x}}^{(t)}$ is defined as in (7.1), $\rho_\tau(u) = u[\tau - I(u<0)]$ is the quantile loss function, and $\sigma > 0$ is a scale parameter.

Obviously, a most natural way to formulate a model based on the approach described in this section is as in Example 24, so that, for each latent state u, we have a specific intercept α_u, whereas the regression parameters $\boldsymbol{\psi}$ are common to all latent states.

7.2.3 Estimation

Given a parametric assumption on the distribution of the response variables, the model log-likelihood has the same expression reported in Section 5.6, that is,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\log f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{x}}_i) = \sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}}\log f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}),$$

where the double sum $\sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}}$ is over all pairs $(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})$ observed at least once. The complete data log-likelihood can be derived as usual:

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) &= \sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}}\sum_{y} a_{u\boldsymbol{x}y}^{(t)}\log f^{(t)}(y|u,\boldsymbol{x})\\ &+ \sum_{u=1}^{k}\sum_{\boldsymbol{x}} b_{u\boldsymbol{x}}^{(1)}\log\pi_{u|\boldsymbol{x}} + \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k}\sum_{\boldsymbol{x}} b_{\bar{u}u\boldsymbol{x}}^{(t)}\log\pi_{u|\bar{u}\boldsymbol{x}}^{(t)},\end{aligned}$$

where, as in Section 5.6, $b_{u\boldsymbol{x}}^{(t)}$ is the frequency of the latent state u and covariate configuration $\boldsymbol{x}$ at occasion t; with reference to the same occasion and covariate configuration, $b_{\bar{u}u\boldsymbol{x}}^{(t)}$ is the number of transitions from state $\bar{u}$ to state u, whereas $a_{u\boldsymbol{x}y}^{(t)}$ is the frequency of subjects that are in latent state u and provide response y.

For what concerns estimation, it is straightforward to check that essentially the same Expectation-Maximization (EM) algorithm described in Section 5.6.1 may be adopted in the present case. In particular, the E-step still consists of computing the expected values of the frequencies involved in the complete data log-likelihood. The M-step still consists of updating the model parameters by maximizing the resulting expected complete data log-likelihood. However, for what concerns the parameters involved in the conditional distribution of the response variable, the M-step is specific for the assumed parametrization. We describe this step in the following.

If a linear regression parametrization, as described in Section 7.2.1, is assumed, the M-step consists of updating the parameter vector $\boldsymbol{\beta}$ by a weighted least square method, that is,

$$\boldsymbol{\beta} = \left(\sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}} \hat{b}_{u\boldsymbol{x}}^{(t)} \boldsymbol{w}_{u\boldsymbol{x}}\boldsymbol{w}_{u\boldsymbol{x}}'\right)^{-1} \sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}}\sum_{y} \hat{a}_{u\boldsymbol{x}y}^{(t)} y \boldsymbol{w}_{ux}^{(t)}. \qquad (7.2)$$

The dispersion parameter σ^2 is instead updated through the following expression:

$$\sigma^2 = \frac{1}{nT}\sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}}\sum_{y} \hat{a}_{uy}^{(t)}(y - \eta_{u\boldsymbol{x}}^{(t)})^2,$$

where $\eta_{u\boldsymbol{x}}^{(t)}$ depends on the value $\boldsymbol{\beta}$ found as above.

Note that, in the absence of individual covariates, expression (7.2) reduces to

$$\alpha_u = \frac{\sum_{t=1}^{T}\sum_{y}\hat{a}_{uy}^{(t)} y}{\sum_{t=1}^{T}\hat{b}_u^{(t)}}, \quad u = 1, \ldots, k;$$

therefore, the conditional expected value of the u-th component corresponds to the weighted average of the outcomes, with suitable weights computed at the E-step.

If a QR parametrization, as described in Section 7.2.2, is instead assumed, at the M-step the parameters $\boldsymbol{\beta}$ are updated by minimizing

$$\sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}}\sum_{y} \hat{a}_{u\boldsymbol{x}y}^{(t)} \rho_\tau(y - \eta_{u\boldsymbol{x}}^{(t)}). \qquad (7.3)$$

This is a weighted minimization of residuals which is commonly performed in the QR literature. Adopting a parametrization as in Example 24, an approach more computationally efficient consists of performing the minimization of (7.3) in two steps, first setting

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{arginf}} \sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}}\sum_{y} \hat{a}_{u\boldsymbol{x}y}^{(t)} \rho_\tau(y - \alpha_u - \boldsymbol{x}'\boldsymbol{\psi}),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$, and then setting

$$\boldsymbol{\psi} = \underset{\boldsymbol{\psi}}{\operatorname{arginf}} \sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}}\sum_{y} \hat{a}_{u\boldsymbol{x}y}^{(t)} \rho_\tau(y - \alpha_u - \boldsymbol{x}'\boldsymbol{\psi}).$$

Each step is a minimization of weighted residuals with an offset, analogously to the third step of the EM algorithm proposed by Geraci and Bottai (2007). A single iteration of these two steps, embedded in the EM algorithm, is sufficient to guarantee convergence of the algorithm. For more details see Farcomeni (2012). Finally, σ is updated as

$$\sigma = \frac{1}{nT}\sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{\boldsymbol{x}}\sum_{y}\hat{a}_{u\boldsymbol{x}y}^{(t)}\rho_{\tau}(y-\eta_{u\boldsymbol{x}}^{(t)}).$$

7.3 Dealing with missing responses

We have so far dealt with balanced panel data in which all subjects are observed at the same number of time occasions T. However, in certain applications this may not happen, leading to a specific number of occasions T_i for every subject i, $i = 1, \dots, n$. This is commonly due to drop out of certain units which leave the panel before the end of the study. In the following, we provide a brief discussion of the types of drop out, referring to the well-known book of Little and Rubin (2002) for a detailed discussion on the topic.

We define drop out to be *completely at random* when the random process generating missingness is independent of the observed and unobserved variables. There is an exogenous independent process generating missingness, and the missing data pattern is randomized and balanced over all covariates and outcomes. This situation is not very common and can be checked heuristically by verifying that approximately the same amount of drop out is observed over all outcome configurations.

We define drop out to be *at random* when missingness depends only on the observed measurements. Consider, for instance, problems of compliance with a treatment, where the outcomes include side effects. The presence of many side effects at the previous occasions can make the patient decide to quit the treatment. If no other factors than the observed outcomes (and covariates) affect the treatment, we have a situation of random drop out. This assumption is rather often reasonable in practice.

We finally define drop out to be *informative* whenever the missingness additionally depends on the outcomes that would have been observed if the subject had not dropped out. Continuing the drug treatment example, suppose a patient decides to quit due, for instance, to sudden side effects which were not observed in the past and do not show up during the visit. This is the same situation as before, with the only difference that the patient drops out one time occasion earlier. We have no information as to why the patient disappeared from the panel. A similar situation is often experienced with responses to questions about income: very high and very low income subjects are more likely to refuse to respond.

The first two cases are known as a situation of *ignorable drop out*, since there is no bias resulting from ignoring the missing data. In the third case *drop out is not ignorable* and pattern mixture or selection models should be specified, including a model for the missing data mechanism.

If we can make a hypothesis of ignorable drop out, we can simply adapt all formulations so far discussed to unbalanced panels, whereas the case of nonignorable drop out requires the formulation of a model for missing data and makes estimation rather difficult. A simple and practical solution, which we just mention, is to impute missing data and work with the completed balanced panel. In principle one can also use an EM algorithm for dealing with missing data.

Concerning ignorable drop out, if subject i definitely leaves the panel after occasion T_i, we can use the same recursion illustrated in Section 3.2 to compute the probability $f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$ for the basic LM model and similar recursions for more complex models. The only difference is that (3.6) is applied for $t = 2, \ldots, T_i$ instead of for $t = 2, \ldots, T$. The same happens for the other recursions which are required, for instance, to compute the posterior probabilities.

A different situation arises with data missing at random before the end of the observation period. For instance, if we observe the response variable only at the first and the third among the available occasions, for the basic LM model we have that

$$f_{\boldsymbol{Y}}((y^{(1)}, y^{(3)})) = \sum_{\boldsymbol{u}} \pi_{u^{(1)}} \pi^{(2)}_{u^{(2)}|u^{(1)}} \pi^{(3)}_{u^{(3)}|u^{(2)}} \phi_{y^{(1)}|u^{(1)}} \phi_{y^{(3)}|u^{(3)}}.$$

In general, modifications of the usual recursions are simple in these cases, even using the matrix notation, as also noted by Zucchini and MacDonald (2009).

As mentioned above, informative drop out is much more difficult to deal with. This topic is rather new in the literature about LM models. We sketch here some ideas which can be useful when drop out may be not ignorable and is due, for instance, to subjects leaving the study because of a terminal event. First of all, with categorical data we can treat the missing values as an additional category for the outcome. Furthermore, we can add a latent state and constrain the $(k+1)$-th latent state to be absorbing. This means that the probability of moving from the latent state to any other is zero, in symbols $\pi_{u|k+1} = 0$ for $u = 1, \ldots, k$. Simultaneously, $\phi_{m|j} = 0$ for $j = 1, \ldots, k$ and $\phi_{m|k+1} = 1$, where m denotes the category "missing" for the outcome. Therefore, no latent state among the first k can yield a missing value, and once subjects move to the $(k+1)$-th latent state they yield missing values until the end of the study. This formulation would allow us to study probability of transitions from the latent states to drop out and other features of the process generating missing values. Furthermore, generalization to the case of informative missing values within the observation period is obtained by simply removing the constraint on the $(k+1)$-th row of the transition matrix.

7.4 Additional computational issues

We detail in this section additional computational issues, some of which have already been mentioned in the previous chapters, and in particular the use of the NR algorithm for maximum likelihood estimation and the use of parametric bootstrap in order to obtain standard errors for the parameter estimates.

7.4.1 Maximization of the likelihood through the Newton-Raphson algorithm

In the following, we describe the NR algorithm in general and then we propose some suggestions about its use for estimating the LM model parameters and give a brief discussion about the related literature.

7.4.1.1 A general description of the algorithm

The NR algorithm is designed to find a root of a differentiable function, via a Taylor's expansion truncated at the first order. It is used in statistics in order to maximize the log-likelihood of a certain model, by finding the root of the first derivative and then the maximum likelihood estimate for an observed sample. The first derivative corresponds to the score vector, usually denoted by $\boldsymbol{s}(\boldsymbol{\theta})$.

More formally, suppose we want to find the root of the score vector $\boldsymbol{s}(\boldsymbol{\theta})$. The Taylor series of $\boldsymbol{s}(\boldsymbol{\theta})$ about $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, truncated at the first order, is given by

$$\boldsymbol{s}(\boldsymbol{\theta}) \approx \boldsymbol{s}(\boldsymbol{\theta}_0) - \boldsymbol{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

where, as usual, $\boldsymbol{J}(\boldsymbol{\theta}_0)$ denotes the observed information matrix, which is equivalent to minus the second derivative of the log-likelihood function. Then, at the first order of approximation, we can find the root of $\boldsymbol{s}(\boldsymbol{\theta})$ as

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1} s(\boldsymbol{\theta}_0). \tag{7.4}$$

The NR algorithm consists of repeating the above step until convergence. The solution at convergence is taken as the maximum likelihood estimate of $\boldsymbol{\theta}$. When computing the observed information matrix is not possible, we may obtain the expected value of $\boldsymbol{J}(\boldsymbol{\theta})$, that is, the Fisher information matrix, and use it in (7.4); the so-called Fisher scoring algorithm results.

As is well known, the NR algorithm can represent a much quicker maximization procedure with respect to the EM algorithm. Moreover, it directly provides an estimate of standard errors, since the observed information matrix is computed within the algorithm. However, the NR algorithm is expected to be much more unstable; in fact, if the starting values are not close to the target, it normally fails to converge for complex models as the LM models.

7.4.1.2 Use for latent Markov models

Even if it is seldom implemented, direct maximization via the NR algorithm can be performed for an LM model. In particular, Lystig and Hughes (2002) and Cappé and Moulines (2005) developed recursive procedures for calculating the derivatives of the log-likelihood for hidden Markov models, which may also be applied for the models of our interest. The availability of the gradient and Hessian makes it possible to directly maximizing the likelihood through the NR algorithm; see also Turner (2008).

According to what was discussed in the previous section, the best use of the NR algorithm in our context is for speeding up the convergence after some iterations of the EM algorithm, which is then intended for finding good starting values. However, in order to overcome some computational issues, it is convenient to reparametrize the model so that the new parameter space has a simpler shape.

Consider, for instance, the basic LM model dealt with in Chapter 3. The parameter space has a complex shape because all parameters are probabilities which, obviously, are constrained to be positive and to sum up to 1 in a suitable way. A more convenient parametrization, at least for performing the NR algorithm, is based on mapping the original parameters on an unbounded space through a series of logits already considered in Chapter 4 to formulate a constrained LM model. In particular, we suggest using reference-category logits for the initial probabilities π_u, $u = 1, \ldots, k$, and for the conditional response probabilities $\phi_{y|u}$, $y = 0, \ldots, c-1$. Moreover, for each row of the transition matrix, we can use logits of the type

$$\log\left(\frac{\pi_{u|\bar{u}}}{\pi_{\overline{u}|\bar{u}}}\right), \quad \bar{u}, u = 1, \ldots, k,\ u \neq \bar{u}.$$

Overall, for the basic LM model we map the original parameters on a space of dimension $\mathbb{R}^{\#\text{par}}$, where #par is defined in (3.1).

Obviously, once the parameters of the reparametrized model have been estimated by the NR algorithm, we can obtain the corresponding estimate of the original parameters by certain simple transformations, which have already been considered in Chapter 4.

7.4.2 Parametric bootstrap

As already mentioned in Sections 2.4 and 3.5.3, an appropriate method to obtain standard errors for the parameter estimates, especially when parameters correspond to probabilities as for the basic LM model, is represented by the parametric bootstrap method. For a general overview on this method see Efron and Tibshirani (1993), Davison and Hinkley (1997), and Chernick (2008). For a discussion in the context of hidden Markov models see Zucchini and MacDonald (2009, Chapter 3). Parametric bootstrap is performed by repeatedly generating data from the estimated model and then fitting the model on the generated data. In the following, this method is illustrated with reference to LM models without covariates.

Let $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimates of the parameters and let $\hat{f}_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$ be the corresponding manifest distribution of the response variables. Then, for $b = 1, \ldots, B$, where B is large enough, the following steps are performed:

1. generate a sample of size n from the distribution $\hat{f}_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$;
2. compute the maximum likelihood estimate of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_b$, on the basis of the data generated above.

In this way, we create the so-called *bootstrap distribution* of $\hat{\boldsymbol{\theta}}$. For each parameter, the standard deviation computed on the basis of this distribution is taken as the standard error to be associated to the corresponding estimate.

The number of bootstrap samples, B, is usually set equal to a large number, such as 999. However, in many cases a number such as $B = 199$ is large enough for a good approximation of the standard errors. Furthermore, there are formal methods which can be used in order to choose the number of bootstrap samples to draw; see Andrews and Buchinsky (2000) among others.

In order to provide an example of the approach described above, consider the basic LM model dealt with in Chapter 3. For this model it is rather simple to draw a sample of size n from the estimated manifest distribution $\hat{f}_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})$. For this aim, in the univariate case we can proceed as follows for $i = 1, \ldots, n$:

TABLE 7.1
Standard errors for the estimates of the conditional response probabilities $\phi_{y|u}$ under model M_5

	$\text{se}(\hat{\phi}_{y\|u})$		
u	$y = 0$	$y = 1$	$y = 2$
1	0.0089	0.0817	0.0182
2	0.0083	0.0813	0.0629
3	0.0031	0.0336	0.0651

TABLE 7.2
Standard errors for the estimates of the initial probabilities π_u under model M_5

u	$\text{se}(\hat{\pi}_u)$
1	0.0244
2	0.0230
3	0.0095

1. draw the latent variable $u_i^{(1)}$ at the first occasion from the distribution $\hat{f}_{U^{(1)}}(u)$ based on the estimated initial probabilities $\hat{\pi}_1, \ldots, \hat{\pi}_k$;

2. for $t = 2, \ldots, T$ draw the latent variable $u_i^{(t)}$ at occasion t from the distribution $\hat{f}_{U^{(t)}|U^{(t-1)}}(u|u_i^{(t-1)})$ based on the estimated transition probabilities $\hat{\pi}_{u|\bar{u}}^{(t)}$;

3. for $t = 1, \ldots, T$ draw $y_i^{(t)}$ from the distribution $\hat{f}_{Y^{(t)}|U^{(t)}}(y|u_i^{(t)})$ based on the conditional response probabilities $\hat{\phi}_{y|u}$.

For an illustration of the above procedure, consider the application of the basic LM model reported in Section 3.7.1, which is based on the marijuana consumption dataset. For these data we obtain the standard errors reported in Tables 7.1, 7.2, and 7.3, which are organized like Tables 3.3, 3.4, and 3.5, respectively.

On the basis of the results in these tables we can assess the level of uncertainty in estimating every model parameter. In particular, we observe that the estimation of the initial probabilities is much more precise than the estimation of the transition probabilities. For instance, the standard error associated with the estimate of π_3 is equal to 0.0095, whereas that for the transition probability $\pi_{3|3}^{(2)}$ is equal to 0.2676. The reason is that a very reduced number of subjects is in the third latent

TABLE 7.3
Standard errors for the estimates of the transition probabilities $\pi^{(t)}_{u|\bar{u}}$ under model M_5

		$\mathrm{se}(\hat{\pi}^{(t)}_{u|\bar{u}})$		
t	$\bar{u}$	$u = 1$	$u = 2$	$u = 3$
2	1	0.0354	0.0360	0.0138
	2	0.1382	0.2030	0.1641
	3	0.1850	0.2237	0.2676
3	1	0.0387	0.0392	0.0167
	2	0.0814	0.1230	0.1011
	3	0.0638	0.1361	0.1471
4	1	0.0412	0.0415	0.0180
	2	0.0695	0.1055	0.0860
	3	0.0375	0.0953	0.0976
5	1	0.0452	0.0441	0.0187
	2	0.0924	0.1196	0.0788
	3	0.0299	0.0881	0.0882

state at the first occasion, and then there is a reduced amount of information on which we can rely to estimate the transition probabilities from this latent state.

7.5 Decoding and forecasting

In the following, we deal with decoding, that is, prediction of the sequence of the latent states for a certain sample unit on the basis of the data observed for this unit. We distinguish between local decoding, which consists of finding the most likely latent state for every time occasion, and global decoding, which consists of finding the most likely sequence of latent states. The second is a more complex problem, which requires suitable algorithms (Viterbi, 1967; Juang and Rabiner, 1991).

In order to present the methods in a simple way, we consider only the case of LM models without covariates. However, the extension to the case of models including covariates may be carried out in a natural way; see Bartolucci and Farcomeni (2009).

In the following, we also deal with forecasting of a future latent state or a future response. Even in this case, in order to keep the presentation simple we explicitly consider the case of a simple LM model without covariates.

TABLE 7.4
Estimated posterior probabilities $\hat{f}_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}})$ under model M_5

	$\hat{f}_{U^{(t)}\mid\tilde{\boldsymbol{Y}}}(u\mid\tilde{\boldsymbol{y}})$		
t	$u=1$	$u=2$	$u=3$
1	0.9850	0.0149	0.0000
2	0.8665	0.1332	0.0003
3	0.0000	0.9957	0.0043
4	0.0000	0.9929	0.0071
5	0.0000	0.9829	0.0171

7.5.1 Local decoding

The local decoding is based on the estimated *posterior probabilities* $\hat{f}_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}})$ which are directly provided by the EM algorithm for $t = 1,\ldots,T$ and $u = 1,\ldots,k$, and every response configuration $\tilde{\boldsymbol{y}}$ observed at least once; see Section 3.5.1 with reference to the basic LM model.

On the basis of the above posterior probabilities, we predict the latent state at occasion t for a sample unit with response configuration $\tilde{\boldsymbol{y}}$, denoted by $\hat{u}^{(t)}(\tilde{\boldsymbol{y}})$, as the value of u which maximizes $\hat{f}_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}})$. Therefore, the entire sequence of predicted latent states produced by the local decoding is $\hat{\boldsymbol{u}}(\tilde{\boldsymbol{y}}) = (\hat{u}^{(1)}(\tilde{\boldsymbol{y}}),\ldots,\hat{u}^{(T)}(\tilde{\boldsymbol{y}}))$.

Provided that the adopted parametrization is based on specific coefficients assigned to every latent state, as in the LM Rasch model described in Example 10, we can also compute the posterior mean of these coefficients through a simple formula of type

$$\hat{\alpha}^{(t)}(\tilde{\boldsymbol{y}}) = \sum_{u=1}^{k} \hat{\alpha}_u \hat{f}_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}).$$

Consider again the version of the basic LM model estimated for the marijuana consumption dataset; see Section 3.7.1. For the sequence of responses $\tilde{\boldsymbol{y}} = (0,0,1,1,1)$, for instance, we have the estimated posterior probabilities reported in the Table 7.4.

Given the estimated posterior probabilities in Table 7.4, for a subject with response configuration $\tilde{\boldsymbol{y}} = (0,0,1,1,1)$, by local decoding we obtain the following sequence of latent states $\hat{\boldsymbol{u}}(\tilde{\boldsymbol{y}}) = (1,1,2,2,2)$, so that there is no difference with respect to the observed sequence considering the different categories for the response variables (from 0 to 2) and for the latent variables (from 1 to 3). On the other hand, for the response configuration $\tilde{\boldsymbol{y}} = (1,2,1,0,0)$ we obtain $\hat{\boldsymbol{u}}(\tilde{\boldsymbol{y}}) = (2,3,2,2,1)$ which is different with respect to the observed sequence in the value predicted at the fourth occasion.

7.5.2 Global decoding

The above approach does not take into account the joint conditional probability of the predicted sequence, given the observed responses. Therefore, when the entire sequence of latent states is of interest, *global decoding* is more appropriate. The predicted sequence of latent states, corresponding to the response vector $\tilde{\boldsymbol{y}}$, is in this case denoted by $\hat{\boldsymbol{u}}^*(\tilde{\boldsymbol{y}}) = (\hat{u}^{*(1)}(\tilde{\boldsymbol{y}}), \ldots, \hat{u}^{*(T)}(\tilde{\boldsymbol{y}}))$. This may be not equal to the sequence $\hat{\boldsymbol{u}}(\tilde{\boldsymbol{y}})$ found by the local decoding method.

In practice, the problem of path prediction described above is tackled by means of the Viterbi algorithm (Viterbi, 1967; Juang and Rabiner, 1991) that we illustrate below. For simplicity we illustrate the algorithm for the basic LM model for univariate response variables.

Let $\hat{p}^{(1)}(u, \tilde{\boldsymbol{y}}) = \hat{f}_{U^{(1)},Y^{(1)}}(u, y^{(1)})$ and, for $t = 2, \ldots, T$, let $\hat{p}^{(t)}(u, \tilde{\boldsymbol{y}})$ be the maximum with respect to $u^{(1)}, \ldots, u^{(t-1)}$ of

$$\hat{f}_{U^{(1)},\ldots,U^{(t-1)},U^{(t)},Y^{(1)},\ldots,Y^{(t)}}(u^{(1)}, \ldots, u^{(t-1)}, u, y^{(1)}, \ldots, y^{(t)}).$$

The Viterbi algorithm is based on a forward recursion to compute the above quantities and a backward recursion for path prediction. In particular, the steps of the algorithm are the following:

1. for $u = 1, \ldots, k$ compute $\hat{p}^{(1)}(u, \tilde{\boldsymbol{y}})$ as $\hat{\pi}_u \hat{\phi}_{y^{(1)}|u}$;
2. for $t = 2, \ldots, T$ and $u = 1, \ldots, k$, compute $\hat{p}^{(t)}(u, \tilde{\boldsymbol{y}})$ as
$$\hat{\phi}_{y^{(t)}|u} \max_{\bar{u}=1,\ldots,k} [\hat{p}^{(t-1)}(\bar{u}, \tilde{\boldsymbol{y}}) \hat{\pi}^{(t)}_{u|\bar{u}}];$$
3. find the optimal state $\hat{u}^{*(T)}(\tilde{\boldsymbol{y}})$ as
$$\hat{u}^{*(T)}(\tilde{\boldsymbol{y}}) = \underset{u=1,\ldots,k}{\operatorname{argmax}}\ \hat{p}^{(T)}(u, \tilde{\boldsymbol{y}});$$
4. for $t = T-1, \ldots, 1$ find $\hat{u}^{*(t)}(\tilde{\boldsymbol{y}})$ as
$$\hat{u}^{*(t)}(\tilde{\boldsymbol{y}}) = \underset{u=1,\ldots,k}{\operatorname{argmax}}\ [\hat{p}^{(t)}(u, \tilde{\boldsymbol{y}}) \hat{\pi}^{(t+1)}_{\hat{u}^{*(t+1)}(\tilde{\boldsymbol{y}})|u}].$$

For the first sequence considered in the previous section, that is, $\tilde{\boldsymbol{y}} = (0, 0, 1, 1, 1)$, we have $\hat{\boldsymbol{u}}^*(\tilde{\boldsymbol{y}}) = (1, 1, 2, 2, 2)$ and then there is no difference with the sequence predicted by local decoding. On the other hand, for the second response configuration previously considered, that is, $\tilde{\boldsymbol{y}} = (1, 2, 1, 0, 0)$, we have $\hat{\boldsymbol{u}}^*(\tilde{\boldsymbol{y}}) = (2, 3, 2, 2, 2)$, which is different for the value assumed at the last occasion, from the sequence predicted by local decoding.

7.5.3 Forecasting

In the following, we deal with forecasting of the latent state and of the response variable at occasion $T+h$, for $h = 1, 2, \ldots$, on the basis of a sequence of responses $\tilde{\boldsymbol{y}}$ and the parameter estimates. In order to keep the presentation simple, we consider an LM model for univariate responses constrained so that the conditional response probabilities and the transition probabilities are time homogeneous.

The situation is more complex when the conditional response probabilities, or the transition probabilities, are time heterogenous and possibly depend on individual covariates. In this case, forecasting requires a sort of *scenario building*, that is, to investigate what would happen for interesting values of time-varying parameters and covariates. However, given its complexity, we skip discussing this issue in detail.

In the simple case mentioned above, with conditional response probabilities which are time homogeneous, as well as the transition probabilities, the latent state at occasion $T+h$ is predicted on the basis of the corresponding *forecasting distribution.* This task is of importance especially when the outcome measures a latent variable with error or when many outcomes at once measure the same latent trait (e.g., health status). The forecasting distribution corresponds to the conditional distribution of $U^{(T+h)}$ given $\tilde{\boldsymbol{Y}}$, that is,

$$\begin{aligned}\hat{f}_{U^{(T+h)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}) &= \frac{\hat{f}_{U^{(T+h)},\tilde{\boldsymbol{Y}}}(u,\tilde{\boldsymbol{y}})}{\hat{f}_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})} \\ &= \frac{\sum_{\bar{u}=1}^{k}\hat{q}^{(T)}(\bar{u},\tilde{\boldsymbol{y}})\hat{f}_{U^{(T+h)}|U^{(T)}}(u|\bar{u})}{\sum_{\bar{u}=1}^{k}\hat{q}^{(T)}(\bar{u},\tilde{\boldsymbol{y}})}, \qquad (7.5)\end{aligned}$$

with $u = 1, \ldots, k$. In the above expression, $\hat{q}^{(T)}(\bar{u},\tilde{\boldsymbol{y}})$ is the estimated joint probability of $U^{(t)} = u$ and $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}}$, see Section 3.2, whereas $\hat{f}_{U^{(T+h)}|U^{(T)}}(u|\bar{u})$ refers to the estimated conditional distribution of $U^{(T+h)}$ given $U^{(T)}$. Both distributions depend on the estimated probabilities $\hat{\phi}_{y|u}$, $\hat{\pi}_u$, and $\hat{\pi}_{u|\bar{u}}$. Once we have computed the probabilities in (7.5), we obtain the predicted latent state at occasion $T+h$, denoted by $\hat{u}^{(T+h)}(\tilde{\boldsymbol{y}})$, as the value of u corresponding to the highest of these probabilities.

It is worth noting that, as the horizon of prediction becomes larger and larger, we obtain the same expression regardless of the observations until occasion T. In particular, we have that

$$\lim_{h\to\infty}\hat{f}_{U^{(T+h)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}) = \hat{\pi}_u^*,$$

where $\hat{\pi}_u^*$ is the estimated stationary probability for the u-th latent state, which depends on the transition probabilities $\hat{\pi}_{u|\bar{u}}$.

Forecasting of the response variable at occasion $T+h$, on the basis of the response configuration $\tilde{\boldsymbol{y}}$ and the parameter estimates, is performed through a forecasting distribution similar to the above one. In particular, we use the distribution

$$\begin{aligned}\hat{f}_{Y^{(T+h)}|\tilde{\boldsymbol{Y}}}(y|\tilde{\boldsymbol{y}}) &= \frac{\hat{f}_{Y^{(T+h)},\tilde{\boldsymbol{Y}}}(y,\tilde{\boldsymbol{y}})}{f_{\tilde{\boldsymbol{Y}}}(\tilde{\boldsymbol{y}})}\\ &= \frac{\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k}\hat{q}^{(T)}(\bar{u},\boldsymbol{y})\hat{f}_{U^{(T+h)}|U^{(T)}}(u|\bar{u})\hat{\phi}_{y|u}}{\sum_{\bar{u}=1}^{k}\hat{q}^{(T)}(\bar{u},\boldsymbol{y})}, \quad (7.6)\end{aligned}$$

with $y = 0,\dots,c-1$, where the same notation as above is adopted. Accordingly, the predicted value of the response variable, denoted by $\hat{y}^{(T+h)}(\tilde{\boldsymbol{y}})$, is obtained by finding the highest among the probabilities in (7.6). Even in this case, it is worth noting that for h tending to infinity we have the forecasting distribution

$$\lim_{h\to\infty}\hat{f}_{Y^{(T+h)}|\tilde{\boldsymbol{Y}}}(y|\tilde{\boldsymbol{y}}) = \sum_{u=1}^{k}\hat{\pi}_u^{*}\hat{\phi}_{y|u},$$

which does not depend anymore on the observed sequence of responses $\tilde{\boldsymbol{y}}$.

7.6 Selection of the number of latent states

As mentioned in the previous chapters, for any adopted LM formulation a crucial point is the choice of the number of latent states, denoted by k, when this number cannot obviously be fixed on the basis of the application of interest. In particular, we suggested the use two information criteria: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC); see Akaike (1973) and Schwarz (1978). These criteria are defined in general in Section 2.5.4; see also Section 3.6 with specific reference to the basic LM model.

We recall that these two criteria are based on suitable indices which jointly measure the goodness-of-fit and the complexity of the model: the lower the value of the index, the better the compromise between goodness-of-fit and complexity. In practice, in order to select the number of latent states by AIC or by BIC, we suggest starting by fitting the model with $k = 1$ and then increasing the number of latent states as long as the corresponding AIC or BIC index decreases. Then, we select the value of k corresponding to the minimum among the computed values of the adopted index.

Given the different definition of the indices involved in AIC and BIC, corresponding to different penalization terms, the two criteria do not always lead to choosing the same number of latent states. In particular, BIC is expected to perform better as the amount of information increases. In the LM literature, the same criteria have been used for model selection by Langeheine (1994), Langeheine and Van de Pol (1994), and Magidson and Vermunt (2001), among many others. However, there is a lively discussion in the literature about the comparative performance of the two criteria; see, for instance, Yang (2005), van Erven et al. (2008), and the references therein.

For all estimation of the above point, we performed a simple simulation study referred to an LM model for bivariate binary responses with time-homogenous conditional response and transition probabilities. This is equivalent to the basic LM model for bivariate responses under the constraint $\pi_{u|\bar{u}}^{(t)} = \pi_{u|\bar{u}}$, $t = 1, \dots, T$. For a more complex simulation study based on an LM with covariates, refer to Bartolucci and Farcomeni (2009).

The simulation study described here considers different scenarios corresponding to different panel lengths ($T = 4, 8$) and sample sizes ($n = 500, 1000$). Moreover, it considers different possibilities for the number of latent states ($k = 1, 2, 3$). Under each scenario, the true conditional response probabilities are fixed as indicated in Table 7.5.

Moreover, for $k = 2, 3$ we use uniform initial probabilities ($\pi_u = 1/k$, $u = 1, \dots, k$) and transition matrix

$$\mathbf{\Pi} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

for $k = 2$ and

$$\mathbf{\Pi} = \begin{pmatrix} 0.80 & 0.15 & 0.05 \\ 0.10 & 0.80 & 0.10 \\ 0.05 & 0.15 & 0.80 \end{pmatrix}$$

for $k = 3$.

TABLE 7.5
Simulation setting about the conditional response probabilities

k	$\boldsymbol{y}$	$\phi_{\boldsymbol{y}\|1}$	$\phi_{\boldsymbol{y}\|2}$	$\phi_{\boldsymbol{y}\|3}$
1	(0,0)	0.2500	-	-
	(0,1)	0.2500	-	-
	(1,0)	0.2500	-	-
	(1,1)	0.2500	-	-
2	(0,0)	0.1137	0.5010	-
	(0,1)	0.1552	0.2301	-
	(1,0)	0.1552	0.2301	-
	(1,1)	0.5758	0.0389	-
3	(0,0)	0.0296	0.2500	0.8488
	(0,1)	0.0462	0.2500	0.0753
	(1,0)	0.0462	0.2500	0.0753
	(1,1)	0.8779	0.2500	0.0005

Under each scenario, we generated 1,000 samples from the true model and for each sample we selected the optimal number of states according to AIC and BIC. In Table 7.6 we report the relative frequency distribution of the predicted k under each considered simulation setting.

We observe that AIC performs generally well as the predicted k is only occasionally different from the true one; when this happens, the former is almost always larger than the latter. On the other hand, BIC has a very good behavior with the exception of certain cases with $n = 500$ and $T = 4$ when, due to a reduced amount of information, it tends to select $k = 2$ latent states.

TABLE 7.6
Predicted number of latent states by AIC and BIC for the bivariate LM model with time-homogenous conditional response and transition probabilities

			Predicted k (AIC)				
T	n	k	1	2	3	4	5
4	500	1	0.908	0.085	0.007	0.000	0.000
		2	0.258	0.697	0.045	0.004	0.000
		3	0.000	0.044	0.856	0.100	0.000
4	1000	1	0.908	0.089	0.003	0.000	0.000
		2	0.009	0.893	0.098	0.003	0.000
		3	0.000	0.000	0.845	0.155	0.003
8	500	1	0.901	0.094	0.005	0.000	0.000
		2	0.021	0.888	0.089	0.002	0.000
		3	0.000	0.000	0.709	0.277	0.014
8	1000	1	0.910	0.087	0.003	0.000	0.000
		2	0.004	0.847	0.144	0.005	0.000
		3	0.000	0.000	0.519	0.454	0.027
			Predicted k (BIC)				
T	n	k	1	2	3	4	5
4	500	1	1.000	0.000	0.000	0.000	0.000
		2	0.000	0.990	0.010	0.000	0.000
		3	0.000	0.938	0.062	0.000	0.000
4	1000	1	1.000	0.000	0.000	0.000	0.000
		2	0.000	1.000	0.000	0.000	0.000
		3	0.000	0.480	0.519	0.000	0.000
8	500	1	1.000	0.000	0.000	0.000	0.000
		2	0.000	0.529	0.471	0.000	0.000
		3	0.000	0.054	0.946	0.000	0.000
8	1000	1	1.000	0.000	0.000	0.000	0.000
		2	0.027	0.973	0.000	0.000	0.000
		3	0.000	0.000	1.000	0.000	0.000

8

Bayesian latent Markov models

8.1 Introduction

In this chapter[1], we consider Bayesian estimation of latent Markov (LM) models as an alternative to maximum likelihood estimation. In the Bayesian framework, specifying appropriate prior distributions on the parameters, which are now random variables, is necessary and also permits the incorporation of prior belief into the model. The posterior distribution of the parameters, which is proportional to the likelihood multiplied by the prior density, is then used for inference on these parameters.

There are different reasons for preferring the Bayesian to the standard likelihood framework, apart from personal beliefs in the Bayesian paradigm. First is availability of prior information: as said, prior information is easily and directly incorporated into the analysis through the prior distributions. Secondly, in the Bayesian framework there is the possibility of working with a posterior distribution on any unknown and, in particular, on the number of latent states k.

In the following, after an illustration of the system of priors for the basic LM model and an extended version of the model with covariates, we describe the Reversible Jump (RJ) Markov chain Monte Carlo (MCMC) algorithm (Green, 1995) for estimating the posterior distribution of the model parameters, which now include the number of latent states. The implementation of the RJ algorithm that we suggest is closely related to that proposed by Robert et al. (2000) and Spezia (2010) for hidden Markov (HM) models. We also outline two alternative algorithms for estimating the posterior distribution: continuous birth and death process (Stephens, 2000; Shi et al., 2002) and parallel sampling (Congdon, 2006). For a more detailed description of the latter, in the context of HM models, see Zucchini and MacDonald (2009). Finally, the Bayesian

[1] Chapter jointly written with Silvia Pandolfi, Department of Economics, Finance, and Statistics, University of Perugia (IT).

approach to LM models is illustrated by an application based on the Panel Survey of Income Dynamics (PSID) data.

8.2 Prior distributions

Prior distributions can be used to summarize available prior information on the parameters. In the absence of prior information, the common approach is to use prior distributions which are justifiable because they lead to special properties for the posterior distribution, or simply because they are convenient. Especially without available prior information, sensitivity to prior choices may be evaluated, for instance, by varying the prior parameters from conservative to optimistic choices and comparing the outcomes (see, for instance, Spiegelhalter et al., 2004). It can be shown that the prior distribution is less and less important as the sample size grows; hence, for large samples its effects are negligible (Bernardo and Smith, 1994). In our experience, latent variable models are often more dependent on prior specification than simpler statistical models are.

In the following, we introduce a system of default priors for the parameters of the basic LM model with univariate responses and then we discuss the case of extended LM models, with particular reference to a model for multivariate data with covariates that will be used in the application to the PSID dataset.

8.2.1 Basic latent Markov model

As illustrated in Chapter 3, the basic LM model for univariate responses is formulated on the basis of the conditional response probabilities

$$\phi_{y|u} = f_{Y^{(t)}|U^{(t)}}(y|u), \quad t = 1,\dots,T,\ u = 1,\dots,k,\ y = 0,\dots,c-1,$$

the initial probabilities of the Markov chain

$$\pi_u = f_{U^{(1)}}(u), \quad u = 1,\dots,k,$$

and the transition probabilities

$$\pi^{(t)}_{u|\bar{u}} = f_{U^{(t)}|U^{(t-1)}}(u|\bar{u}), \quad t = 2,\dots,T,\ \bar{u}, u = 1,\dots,k.$$

In specifying default prior distributions for the above probabilities we adopt the approach of Cappé et al. (2005) and Spezia (2010), who employ a transformation based on unnormalized probabilities, which facilitates the implementation of the RJ estimation algorithm that will be illustrated in Section 8.3.1.

Let $Ga(a, b)$ denote the gamma distribution with parameters a and b. We express the conditional response probabilities as

$$\phi_{y|u} = \frac{\tilde{\phi}_{y|u}}{\sum_{h=0}^{c-1} \tilde{\phi}_{h|u}}, \quad u = 1, \dots, k,\ y = 0, \dots, c-1,$$

where $\tilde{\phi}_{y|u}$ are unnormalized probabilities, which are assumed to be *a priori* independent and with distribution $\tilde{\phi}_{y|u} \sim Ga(\delta_{uy}, 1)$. As default prior we set $\delta_{uy} = 1$ for all u and y. This choice is equivalent to assuming that each vector $\boldsymbol{\phi}_u$ with elements $\phi_{y|u}$, $y = 0, \dots, c-1$, has a Dirichlet prior distribution with parameter equal to a vector of ones, that is $\mathbf{1}_c$; this is the default prior for multinomial cell probabilities (Tuyl et al., 2009).

Similarly, regarding the other parameters, we let

$$\pi_u = \frac{\tilde{\pi}_u}{\sum_{v=1}^{k} \tilde{\pi}_v}, \quad u = 1, \dots, k, \tag{8.1}$$

where $\tilde{\pi}_u$ are assumed *a priori* independent, with distribution $Ga(\delta_u, 1)$ for $u = 1, \dots, k$. Moreover, we let

$$\pi_{u|\bar{u}}^{(t)} = \frac{\tilde{\pi}_{u|\bar{u}}^{(t)}}{\sum_{v=1}^{k} \tilde{\pi}_{v|\bar{u}}^{(t)}}, \quad t = 2, \dots, T,\ \bar{u}, u = 1, \dots, k, \tag{8.2}$$

where $\tilde{\pi}_{u|\bar{u}}^{(t)}$ are assumed *a priori* independent with distribution $Ga(\delta_{\bar{u}u}^{(t)}, 1)$ for all t, $\bar{u}$, and u. Again, we set the default hyperparameters as $\delta_u = \delta_{\bar{u}u}^{(t)} = 1$, yielding a Dirichlet prior distribution with parameter $\mathbf{1}_k$ for the vector $\boldsymbol{\pi}$ with elements π_u, $u = 1, \dots, k$, and for each row of the transition probability matrix with elements $\pi_{u|\bar{u}}^{(t)}$, $\bar{u}$, $u = 1, \dots, k$. Another possible choice regarding the transition probabilities is

$$\delta_{\bar{u}u}^{(t)} = kI(\bar{u} = u) + sI(\bar{u} \neq u), \quad t = 2, \dots, T,\ \bar{u}, u = 1, \dots, k, \tag{8.3}$$

for a suitable constant s which is typically much smaller than k. In this way we express the hypothesis that the probability of persistence is greater than the other probabilities in every transition matrix (Spezia, 2010).

Finally, for the parameter k we define a discrete uniform prior distribution between 1 and $k_{\max}$, where $k_{\max}$ is the maximum number of latent states we *a priori* admit. Usually $k_{\max}$ is greater than the number of latent states of the most complex model likely to be visited by the estimation algorithm.

8.2.2 Constrained and extended latent Markov models

As shown in Chapter 4, constraints on the measurement model or on the latent model are typically expressed on the basis of a formulation in which a block of probabilities (conditional response, initial, and transition) is parametrized on the basis of a vector of parameters contained into a space of type $\mathbb{R}^h$ for a suitable h. This happens, for instance, for the LM Rasch model described in Example 10, in which the conditional response probabilities depend on the parameters α_u and $\psi^{(t)}$.

When a block of probabilities of the model is substituted by parameters belonging to a space of type $\mathbb{R}^h$, we simply assume *a priori* that these parameters follow a zero-centered multivariate Gaussian distribution, with covariance matrix either diagonal or proportional to the inverse of the variance-covariance matrix of the covariates (Leonard, 1975; Nazaret, 1987). For the probabilities that are not parametrized, and that directly correspond to model parameters, we assume prior distributions of the type described in Section 8.2.1. Similarly, for the number of latent states k we still assume a uniform prior distribution from 1 to a certain $k_{\max}$.

Even in complex cases, an extended LM model is typically based on parameters included in a space of type $\mathbb{R}^h$. Consequently, we formulate the prior distribution on the model parameters as described above. In this regard, it is useful to describe in more detail an example, which is related to Example 18 and is the basis for the application to the PSID data.

Example 25 — Marginal model for bivariate binary responses. *Consider an LM model with covariates for bivariate binary responses. We assume the parametrization of the conditional response probabilities based on the following assumptions:*

$$\begin{aligned}
\log\frac{\phi^{(t)}_{j1|u\boldsymbol{x}}}{\phi^{(t)}_{j0|u\boldsymbol{x}}} &= \alpha_{ju}+\boldsymbol{x}'\boldsymbol{\psi}_j, \quad j=1,2,\\
\log\frac{\phi^{(t)}_{(0,0)|u\boldsymbol{x}}\phi^{(t)}_{(1,1)|u\boldsymbol{x}}}{\phi^{(t)}_{(0,1)|u\boldsymbol{x}}\phi^{(t)}_{(1,0)|u\boldsymbol{x}}} &= \psi_3.
\end{aligned}$$

Note that α_{1u} and $\boldsymbol{\psi}_1$ affect the distribution of the first response variable (marginal with respect to the second variable), α_{2u} and $\boldsymbol{\psi}_2$ affect the distribution of the second response variable (marginal with respect to the first variable), and ψ_3 measures the dependence between these two variables. Moreover, we assume that the transition probabilities of the Markov chain are time homogeneous and independent of the covariates; similarly, the initial probabilities are assumed to be independent of the *covariates. Therefore, further to the above parameters, we have the parameters* π_u *and* $\pi_{u|\bar{u}}$ *in the model. These are expressed in terms of the unnormalized probabilities* $\tilde{\pi}_u$ *and* $\tilde{\pi}_{u|\bar{u}}$*, as clarified in (8.1) and (8.2).*

For the LM model based on the these assumptions, we assume that all parameters are a priori *independent. Moreover, the prior distributions are specified as follows:*

1. $N(0, \sigma_\alpha^2)$ *for* α_{1u} *and* α_{2u}*,*
2. $N(0, \sigma_\psi^2)$ *for every element of* $\boldsymbol{\psi}_1$ *and* $\boldsymbol{\psi}_2$ *and for* ψ_3*,*
3. $Ga(1, 1)$ *for every unnormalized probability* $\tilde{\pi}_u$*,*
4. $Ga(\delta_{\bar{u}u}, 1)$ *for every unnormalized probability* $\tilde{\pi}_{\bar{u}|u}$*, with* $\delta_{\bar{u}u}$ *specified as in (8.3).*

8.3 Bayesian inference via Reversible Jump

Bayesian inference is performed by the posterior distribution of the model parameters, which depends on the prior specification and the observed sample of data. It is rather obvious that for LM models we have not an explicit expression for the posterior distribution of the parameters. The typical solution to this problem consists of using a Monte Carlo algorithm, which allows us to draw a sample of arbitrary dimension from the posterior distribution. In particular, we suggest to use the RJ MCMC method proposed by Green (1995). This algorithm explores models with a different number of latent states, by jumping from one model to another. As for the other parameters, the number of states is then estimated on the basis of its simulated posterior distribution.

In the following we illustrate the structure of the RJ algorithm. We also deal with the label-switching problem, and then we describe how to suitably use the RJ output for Bayesian inference on the assumed LM model for the data at hand.

8.3.1 Reversible Jump algorithm

The implementation of the RJ algorithm is based on two different types of move. Moves of the first type are aimed at updating the parameters of the model given the current number of states k; moves of the second type are aimed at updating the number of states.

We now outline these types of move and, later, we give a more detailed explanation of each of them:

- **Metropolis-Hastings (MH) move**: given the current k, this move consists of drawing the model parameters from their posterior distribution.

- **Split/combine moves** (each proposed with probability 0.5): in the split move, a state is chosen at random and then it is split into two new states so as to increase the current value of k by one; in the combine move, a pair of states is chosen at random and merged into a new one, so as to reduce k by one.

- **Birth/death moves** (each proposed with probability 0.5): the birth move is accomplished by generating a new state and drawing the new parameters from their respective priors; in the death move a state is selected at random and then deleted along with the corresponding parameters. Similar to split/combine moves, birth/death moves allow us to increase or decrease the current k by one.

This structure recalls that of the RJ algorithm for mixture models proposed by Richardson and Green (1997), although the birth/death moves are not limited to empty components; see, among others, Cappé et al. (2005, Chapter 13) and Spezia (2010). Furthermore, instead of passing through each move deterministically, at every step of the RJ algorithm we first perform an MH move and then we randomly choose, with probability equal to 0.5, to perform a split/combine or a birth/death move. Finally, our implementation does not simulate the latent process since it directly uses the manifest (or marginal) distribution, which is computed by the usual recursion of Baum et al. (1970), as clarified in the previous chapters. This is because the first type of move is based on an MH sampler rather than on a Gibbs sampler as in the method proposed by Richardson and Green (1997).

To improve the convergence and the mixing of the algorithm, we also apply the *random permutation sampling* of Frühwirth-Schnatter (2001). We do not select *a priori* any artificial identifiability constraints, and the posterior density is left unconstrained. The algorithm is then free to visit all the possible $k!$ subspaces obtained by permuting the labeling of the k latent state. To force the algorithm to visit all subspaces on which the posterior density is defined about the same number of times, we conclude any sweep of the algorithm by permuting the generated values according to the random selection of one of the possible $k!$ ordering of the states.

In detail, we illustrate the moves outlined above; regarding the notation, we use k for the current number of states and $\boldsymbol{\theta}_k$ for the current parameter vector. Moreover, we denote the model likelihood by $L(\boldsymbol{\theta}_k)$ and by $f(\boldsymbol{\theta}_k)$ the density of the prior distribution on these parameters.

Metropolis-Hastings move

Retaining the current k, we draw model parameters from their posterior distribution by a series of multiplicative random walk moves. Moreover, we consider a logarithmic transformation of the unnormalized probabilities $\tilde{\phi}_{y|u}$, $\tilde{\pi}_u$, and $\tilde{\pi}_{u|\bar{u}}^{(t)}$. In particular, when dealing with the basic LM model, these parameters are updated as follows:

1. $\log \tilde{\phi}_{y|u}^* = \log \tilde{\phi}_{y|u} + \varepsilon_{y|u}$, with $u = 1, \ldots, k$, $y = 0, \ldots, c-1$;
2. $\log \tilde{\pi}_u^* = \log \tilde{\pi}_u + \varepsilon_u$, with $u = 1, \ldots, k$;
3. $\log \tilde{\pi}_{u|\bar{u}}^{*(t)} = \log \tilde{\pi}_{u|\bar{u}}^{(t)} + \varepsilon_{u|\bar{u}}^{(t)}$, with $t = 2, \ldots, T$, $\bar{u}, u = 1, \ldots, k$.

In the above expressions, every random term $\varepsilon_{y|u}$, ε_u, and $\varepsilon_{u|\bar{u}}^{(t)}$ is drawn from a normal distribution centered at 0 and with suitable variance. Moreover, according to standard rules (Metropolis et al., 1953; Hastings, 1970) the proposed parameters are accepted with probability equal to $\min\{1, A\}$ with

$$A = \frac{L(\boldsymbol{\theta}_k^*)}{L(\boldsymbol{\theta}_k)} \frac{f(\boldsymbol{\theta}_k^*)}{f(\boldsymbol{\theta}_k)} |J_{\mathrm{MH}}|,$$

where $\boldsymbol{\theta}_k^*$ is the proposed vector of parameters drawn as above and J_{MH} is the Jacobian that arises because we work with a log-scale transformation. Note that, in order to increase the acceptance rate, each block of parameters may be updated by a separate MH move.

The above moves are slightly different when we deal with extended LM models. For instance, if the model of interest is the one described in Example 25, the first proposal, which is used for the parameters $\tilde{\phi}_{y|u}^*$, is substituted with

1a. $\alpha_{ju}^* = \alpha_{ju} + \varepsilon_{ju}$;

1b. $\boldsymbol{\psi}_j^* = \boldsymbol{\psi}_j + \boldsymbol{\varepsilon}_j$, with $j = 1, 2$;

1c. $\psi_3^* = \psi_3 + \varepsilon$.

Every random variable ε_{ju} and ε and those in $\boldsymbol{\varepsilon}_j$ are independent and have a normal distribution with suitable variance. Moreover, the last proposals, concerning the parameters $\tilde{\pi}_u$ and $\tilde{\pi}_{u|\bar{u}}$, are formulated as above.

Split/combine moves

In the split move a state u_0 is randomly selected and split it into two new ones, u_1 and u_2. For the basic LM model, the corresponding parameters are split as follows:

1. $\tilde{\phi}^*_{y|u_1} = \tilde{\phi}_{y|u_0}\xi_y$, $\tilde{\phi}^*_{y|u_2} = \tilde{\phi}_{y|u_0}/\xi_y$, with $y = 0, \dots, c-1$;

2. $\tilde{\pi}^*_{u_1} = \tilde{\pi}_{u_0}\zeta$, $\tilde{\pi}^*_{u_2} = \tilde{\pi}_{u_0}(1-\zeta)$;

3a. $\tilde{\pi}^{*(t)}_{u_1|\bar{u}} = \tilde{\pi}^{(t)}_{u_0|\bar{u}}\zeta^{(t)}_{\bar{u}}$, $\tilde{\pi}^{*(t)}_{u_2|\bar{u}} = \tilde{\pi}^{(t)}_{u_0|\bar{u}}(1-\zeta^{(t)}_{\bar{u}})$, with $t = 2, \dots, T$, $\bar{u} = 1, \dots, k$, $\bar{u} \neq u_0$;

3b. $\tilde{\pi}^{*(t)}_{u|u_1} = \tilde{\pi}^{(t)}_{u|u_0}\xi^{(t)}_u$, $\tilde{\pi}^{*(t)}_{u|u_2} = \tilde{\pi}^{(t)}_{u|u_0}/\xi^{(t)}_u$, with $t = 2, \dots, T$, $u = 1, \dots, k$, $u \neq u_0$;

3c. Split $\tilde{\pi}^{(t)}_{u_0|u_0}$ proposing

$$\begin{aligned}
\tilde{\pi}^{*(t)}_{u_1|u_1} &= \tilde{\pi}^{(t)}_{u_0|u_0}\zeta^{(t)}\omega^{(t)}_1, \quad \tilde{\pi}^{*(t)}_{u_2|u_1} = \tilde{\pi}^{(t)}_{u_0|u_0}(1-\zeta^{(t)})\omega^{(t)}_2,\\
\tilde{\pi}^{*(t)}_{u_1|u_2} &= \tilde{\pi}^{(t)}_{u_0|u_0}\zeta^{(t)}/\omega^{(t)}_1, \quad \tilde{\pi}^{*(t)}_{u_2|u_2} = \tilde{\pi}^{(t)}_{u_0|u_0}(1-\zeta^{(t)})/\omega^{(t)}_2,
\end{aligned}$$

with $t = 2, \dots, T$.

Every error term ξ_y, $\xi^{(t)}_u$, $\omega^{(t)}_1$, and $\omega^{(t)}_2$ has a gamma distribution with suitable parameters, whereas every error term ζ, $\zeta^{(t)}_{\bar{u}}$, and $\zeta^{(t)}$ has a uniform distribution between 0 and 1. The most complex proposal is the third, which is for splitting the elements in the u_0-th column of every transition matrix, for splitting the elements in the u_0-th row, and for splitting the diagonal element $\tilde{\pi}^{(t)}_{u_0|u_0}$.

The proposed parameter values are accepted with probability $\min\{1, A\}$, with

$$\begin{aligned}
A &= \frac{L(\boldsymbol{\theta}^*_{k+1})}{L(\boldsymbol{\theta}_k)}\frac{f(\boldsymbol{\theta}^*_{k+1})}{f(\boldsymbol{\theta}_k)}\frac{p_{\mathrm{c}}(k+1)}{p_{\mathrm{s}}(k)}\\
&\times \frac{|J_{\mathrm{split}}|}{\left[\prod_{y=0}^{c-1} g(\xi_y)\right]\left[\prod_{t=2}^{T}\prod_{u\neq u_0} g(\xi^{(t)}_u)\right]\left[\prod_{t=2}^{T} g(\omega^{(t)}_1)g(\omega^{(t)}_2)\right]},
\end{aligned} \quad (8.4)$$

where $p_{\mathrm{c}}(k+1)$ is the probability of performing a combine move when the current number of states is $k+1$, $p_{\mathrm{s}}(k)$ is the probability of performing a split move when the current number of states is k, J_{split} stands for the Jacobian of the transformation from $\boldsymbol{\theta}_k$ to $\boldsymbol{\theta}^*_{k+1}$, and $g(\cdot)$ stands for the density of a proposal distribution. Moreover, the proposals of the random variables with uniform distribution, that is, ζ, $\zeta^{(t)}_{\bar{u}}$, and $\zeta^{(t)}$, do not explicitly appear because they are always equal to 1.

Finally, when we are dealing with the model in Example 25, which includes individual covariates, the split move is performed essentially in the same way, apart from updating the parameters in the measurement model as follows:

1. $\alpha^*_{ju_1} = \alpha_{ju_0} - \varepsilon_j, \quad \alpha^*_{ju_2} = \alpha_{ju_0} + \varepsilon_j$, with $j = 1, 2$,

where every ε_j has a normal distribution with mean 0 and suitable variance. The other proposals are essentially as above.

In the reverse combine move two distinct states, u_1 and u_2, are chosen at random and merged into a single state u_0, so as to preserve reversibility. For the basic LM model, we have to perform the following operations:

1. $\tilde{\phi}^*_{y|u_0} = \sqrt{\tilde{\phi}_{y|u_1}\tilde{\phi}_{y|u_2}}$, with $y = 0, \ldots, c-1$;

2. $\tilde{\pi}^*_{u_0} = \tilde{\pi}_{u_1} + \tilde{\pi}_{u_2}$;

3a. $\tilde{\pi}^{*(t)}_{u_0|\bar{u}} = \tilde{\pi}^{(t)}_{u_1|\bar{u}} + \tilde{\pi}^{(t)}_{u_2|\bar{u}}$, with $t = 2, \ldots, T$, $\bar{u} = 1, \ldots, k$, $\bar{u} \neq u_0$;

3b. $\tilde{\pi}^{*(t)}_{u|u_0} = \sqrt{\tilde{\pi}^{(t)}_{u|u_1}\tilde{\pi}^{(t)}_{u|u_2}}$, with $t = 2, \ldots, T$, $u = 1, \ldots, k$, $u \neq u_0$;

3c. $\tilde{\pi}^{*(t)}_{u_0|u_0} = \sqrt{\tilde{\pi}^{(t)}_{u_1|u_1}\tilde{\pi}^{(t)}_{u_1|u_2}} + \sqrt{\tilde{\pi}^{(t)}_{u_2|u_1}\tilde{\pi}^{(t)}_{u_2|u_2}}$, with $t = 2, \ldots, T$.

The combine move is accepted with probability $\min\{1, A^{-1}\}$, where A is defined as in (8.4). Moreover, for the model in Example 25, the combine move consists of performing:

1. $\alpha_{ju_0} = (\alpha_{ju_1} + \alpha_{ju_2})/2$, with $j = 1, 2$.

The other proposals are essentially as above.

Note that the split/combine moves do not influence the parameters $\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$, and ψ_3 as they are not affected by the number of states.

Birth/death moves

The birth move is accomplished by generating a new state, denoted by u_0, drawing the new parameters from their respective priors. The remaining parameters are simply copied to the proposed new parameter vector $\boldsymbol{\theta}^*_{k+1}$. In the death move a state u_0 is selected at random and then deleted along with the corresponding parameters.

The acceptance probability of the birth move is $\min\{1, A\}$, where A is simply

$$A = \frac{L(\boldsymbol{\theta}^*_{k+1})}{L(\boldsymbol{\theta}_k)} \frac{p_{\mathrm{d}}(k+1)}{p_{\mathrm{b}}(k)},$$

where $p_{\mathrm{d}}(k+1)$ is the probability of performing a death move when the current number of states is $k+1$ and $p_{\mathrm{b}}(k)$ is the probability of performing a birth move when the current number of states is k. The death move is accepted with probability $\min\{1, A^{-1}\}$.

The simplicity of the expression for A given above is because the proposal densities are equal to the prior densities of the corresponding parameters and because the components in $\boldsymbol{\theta}_k$ remain the same in $\boldsymbol{\theta}_{k+1}$ and the determinant of the Jacobian of the transformation is in this case equal to 1.

8.3.2 Post-processing the Reversible Jump output

Since we use labeling-invariant priors and we do not put constraints on the RJ steps, the label-switching problem, already described in Section 3.4, arises. We recall that this problem is related to the invariability of the model likelihood if the parameters associated with two or more different states are exchanged.

In order to face with the label switching problem, we post-process the output of the RJ algorithm adopting essentially the same approach as Marin et al. (2005). In summary, we sort the latent states at the end of every RJ iteration on the basis of the permutation of these states which minimizes the distance from the posterior mode. More details are given in the following. Note that alternative methods have been proposed in the literature to deal with the label switching problem, such as the artificial identifiability constraints method (Diebolt and Robert, 1994; Richardson and Green, 1997) and the related random permutation sampling (Frühwirth-Schnatter, 2001), the relabeling algorithm (Celeux, 1998; Stephens, 2000), and the label-invariant loss function method (Celeux et al., 2000; Hurn et al., 2003); for general review, see Jasra et al. (2005) and Spezia (2009).

Following Marin et al. (2005), our method of post-processing consists of selecting the model with the number of states $\hat{k}$ that has been visited most often by the RJ algorithm, after discarding a suitable number of iterations as a burn-in. Let $\boldsymbol{\theta}^{(1)}_{\hat{k}}, \ldots, \boldsymbol{\theta}^{(m_{\hat{k}})}_{\hat{k}}$ denote the sequence of random draws from this model, where, in general, m_k is the number of times that the RJ algorithm visited the model with k states. Then, the post-processing proceeds as follows:

1. compute the posterior mode $\hat{\boldsymbol{\theta}}_{\hat{k}}$ under the selected model as

$$\hat{\boldsymbol{\theta}}_{\hat{k}} = \underset{h=1,\ldots,m_{\hat{k}}}{\operatorname{argmax}} \; L(\boldsymbol{\theta}_{\hat{k}}^{(h)}) f(\boldsymbol{\theta}_{\hat{k}}^{(h)});$$

2. let $\text{perm}(\boldsymbol{\theta}_{\hat{k}}^{(h)})$ denote the vector $\boldsymbol{\theta}_{\hat{k}}^{(h)}$ with elements permuted according to a certain permutation $\text{perm}(\cdot)$ of the latent states and let $\mathcal{P}$ denote the space of all possible permutations;

3. for $h = 1, \ldots, m_{\hat{k}}$, substitute $\boldsymbol{\theta}_{\hat{k}}^{(h)}$ with the corresponding permutation which minimizes the distance from $\hat{\boldsymbol{\theta}}_{\hat{k}}$, that is,

$$\underset{\text{perm}(\cdot)\in\mathcal{P}}{\operatorname{argmin}} \; \|\text{perm}(\boldsymbol{\theta}_{\hat{k}}^{(h)}) - \hat{\boldsymbol{\theta}}_{\hat{k}}\|,$$

   where $\| \cdot \|$ denote the norm operator.

8.3.3 Inference based on the simulated posterior distribution

Once the RJ algorithm has been performed for a number of iterations that ensures the convergence to the posterior distribution of the model parameters, and after that the initial draws are eliminated as a burn-in, the remaining draws can be used for inference.

First of all, the posterior distribution of the number of latent states, k, is estimated by the corresponding frequency distribution based on the RJ draws, that is, on the basis of the frequencies m_k, $k = 1, \ldots, k_{\max}$, suitably rescaled. Then, if this distribution is very concentrated on a particular value of k, we have little uncertainty related to the number of latent states whose estimate, as in the previous section, is indicated by $\hat{k}$. Otherwise, we can take into account different values of k in a logic of multimodel inference. Moreover, in order to improve the efficiency of the estimation of the posterior distribution of k, we can use the approach Bartolucci et al. (2006). In addition, consider that, by using the estimated posterior distribution at issue, we can simply obtain the Bayes factor between two competing models, corresponding to different values of k, which is interpreted as suggested by Kass and Raftery (1995). The Bayes factor is strongly related to the Bayesian information criterion (Schwarz, 1978), which in the previous chapters has been suggested as selection criterion in connection to standard likelihood inference; see, in particular, Sections 2.5.4 and 3.6.

Once the number of states, $\hat{k}$, has been selected, the model parameters are estimated by the average or the mode (as in Section 8.3.2) of the corresponding simulated posterior distribution, that is, on the basis of the RJ draws $\boldsymbol{\theta}_{\hat{k}}^{(h)}$, $h = 1, \ldots, m_{\hat{k}}$. Credible intervals can be obtained through the corresponding quantiles, for instance through the 2.5-th and the 97.5-th percentiles. Finally, even if more formal procedures exist, a widespread practice for testing hypotheses in the context of Bayesian inference is based on inverting a credible interval and then rejecting the hypothesis $H_0 : \theta = \theta_0$ if θ_0 is not included in such an interval, where θ denotes a generic model parameter.

8.4 Alternative sampling strategy

The RJ algorithm illustrated in the previous section is convenient in that a reasonable mixing is often obtained with respect to the parameter k. On the other hand, the necessity of obtaining a convenient expression for the Jacobians involved in the acceptance probabilities force us to use Metropolis random walks on the other parameters, which need a fine tuning. Therefore, in the following we outline two alternative sampling algorithms.

8.4.1 Continuous birth and death process based on data augmentation

We propose in this section an alternative sampling strategy which is based on a continuous birth and death (BD) MCMC approach (Stephens, 2000; Shi et al., 2002). This strategy is based on two key ideas. First of all, the likelihood is augmented with the unknown latent state indicators, denote by $u_i^{(t)}$, $i = 1, \ldots, n$, $t = 1, \ldots, T$, basically working with the complete data likelihood also used in the Expectation-Maximization algorithm for maximum likelihood estimation. Moreover, transdimensional moves, which allow us to change the number of states, are based on a continuous BD process. In most cases, this approach allows us to sample all parameters related to the latent process from their full conditionals, except from the parameter k. In addition, a sample from the posterior of the latent indicators is obtained as a natural by-product. We found this sampling strategy a computationally efficient strategy, even if it is less known than the RJ algorithm.

The algorithm at issue performs a series of moves that are summarized in the following. The move that clearly distinguishes the present algorithm from the RJ algorithm is the last one.

1. **Reorder of the latent states**: this consists of reordering the

latent states according to some criteria, such as that the probability of "success" increases with the latent state indicator, in the case of response binary variables. By fixing an order *a priori*, we are implicitly choosing one of the $k!$ modes of the likelihood that could be visited by the algorithm and excluding all the others. Frühwirth-Schnatter (2001) discussed in detail the advantages and disadvantages of this strategy, which may be considered as a simple solution to the label switching problem, alternative to the strategy discussed in Section 8.3.2.

2. **Update of the latent indicators**: the latent indicators $u_i^{(t)}$ are updated by the *ff-bs* algorithm proposed by Chib (1996). This consists of sampling, for $i = 1, \ldots, n$, the indicator $u_i^{(T)}$ from the full conditional distribution given the current value of parameters and then $u_i^{(t)}$ for $t = 1, \ldots, T-1$ in reverse order. Note that each full conditional is a multinomial distribution with frequency 1 and probability vector that may be obtained through the usual forward recursion of Baum et al. (1970).

3. **Update of the model parameters**: under the basic LM model, the system of priors defined in Section 8.2.1 is equivalent to a system based on Dirichlet prior distributions. In this case, due to the use of a data augmentation scheme, we can update every vector of probabilities (conditional response, initial, transition) from the corresponding full conditional distribution which still belongs to the Dirichlet family. In practice, we are adopting a Gibbs sampler to update these parameters. On the other hand, in case of an extended LM model with covariates, the full conditional for the parameters involved in the manifest distribution is not available in closed form, and then, for certain blocks of parameters, we need to rely on an MH sampler of the type described in Section 8.3.1.

4. **Continuous BD**: this is aimed at updating the parameter k. The continuous process is started from the current value of k and run for a fixed time. The value of k is maintained for a time which is drawn from an exponential distribution with parameter depending on two quantities called *birth rate* and *death rate*. At the end of the time drawn as above, we randomly choose between a birth and a death move with probability again depending on the these two rates. These moves are performed in a way similar to the one described in Section 8.3.1 for the RJ algorithm, but they are accepted with probability 1.

Obviously, in order to make proper inference on the basis of the output of the continuous BD algorithm at issue, we have to properly take into account the permanence time of the algorithm in every visited state.

8.4.2 Parallel sampling

Congdon (2006) proposed a very simple model selection strategy based on parallel sampling. The approach uses a Monte Carlo approximation of the posterior distribution for each possible model, namely for $k = 1, \dots, k_{\max}$. This operation can be very time consuming unless $k_{\max}$ is small, that is, unless we are sure that the number of latent states is not larger than a certain small $k_{\max}$.

In practice, for any fixed k, we perform a fixed number m_k of iterations. At each iteration h we draw $\boldsymbol{\theta}_k^{(h)}$ from the posterior distribution of the model parameters by a standard MCMC algorithm. Then we compute the posterior probability of k as the sample average of the m_k corresponding posterior probabilities conditional also on $\boldsymbol{\theta}_k^{(h)}$. See Zucchini and MacDonald (2009) for a detailed illustration of this algorithm for HM models.

8.5 Application to the labor market dataset

In this section, we illustrate the application of the RJ algorithm, introduced in Section 8.3.1, to estimate the LM model with covariates and bivariate responses described in Example 25 for the dataset deriving from the PSID; see Section 1.4.3. With respect to the application illustrated in Section 5.10.2, we restrict the analysis to a sample of $n = 482$ women, selecting one record out of every three. We refer to a smaller sample so as to perform a larger number of iterations of the RJ algorithm and, consequently, to efficiently approximate the posterior distribution of the model parameters. Note that, with respect to the model considered in Section 5.10.2, the model estimated here does not include the lagged responses among the covariates entering in the expressions for the logits referring to the conditional distribution of each response variable.

For the above model, we used the prior distributions defined in Example 25. In particular, we considered two values of the variance of the zero-centered Gaussian distribution for α_{1u} and α_{2u}, that is, $\sigma_\alpha^2 = 5, 10$. The smaller value of this variance expresses the prior belief that the parameters α_{1u} and α_{2u}, associated with the different latent states, are closer among the latent states. This implies a less significant unobserved heterogeneity between subjects with respect to a choice of the prior vari-

TABLE 8.1
Posterior probabilities of the number of latent states for $\sigma^2_\alpha = \sigma^2_\psi = 5, 10$

k	$\sigma^2_\alpha = \sigma^2_\psi = 5$	$\sigma^2_\alpha = \sigma^2_\psi = 10$
≤ 3	0.081	0.006
4	0.554	0.298
5	0.320	0.580
6	0.044	0.107
≥ 7	0.003	0.009

ance of $\sigma^2_\alpha = 10$. The same two values are also considered for the variance of the Gaussian prior distribution for the parameters in $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ and for ψ_3; that is, we let $\sigma^2_\psi = 5, 10$.

The parameters used for the proposal distributions in the MH move, performed with fixed k, were tuned so as to obtain acceptance rates in the range of 0.15–0.30. A similar strategy was used to choose the parameters of the gamma distribution for the error terms in the split/combine move. Moreover, the RJ algorithm was initialized from the maximum likelihood estimate obtained with $k = 1$. Then, we ran the algorithm for 1000000 iterations discarding the first 200000 iterations as a burn-in.

On the basis of the RJ output we obtained the estimated posterior probabilities for the number of states, k, which are reported in Table 8.1. We observe that the algorithm leads to choosing a model with a number of states between 4 and 5, on the basis of the values of the prior parameters. In particular, as clarified above, the choice of a smaller number of states is consistent with a lower *a priori* dispersion for the parameters α_{1u} and α_{2u}. The sensitivity of the inferential results to the prior specification represents an aspect that require a separated and more detailed analysis.

In Table 8.2 we report the acceptance rates for the different moves. Due to the complexity of the model and the very large dimension of the parameter space, these acceptance rates are low for the transdimensional moves, especially for the split/combine moves.

In Table 8.3 we show the estimates of the parameters collected in the vectors $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ and of the parameter ψ_3, for $\sigma^2_\alpha = \sigma^2_\psi = 5$ and $k = 4$ and for $\sigma^2_\alpha = \sigma^2_\psi = 10$ and $k = 5$. In Table 8.4 we also illustrate the same estimates computed by the average over all the draws (after discarding the burn-in) without limiting this average to the draws corresponding to the selected number of states. These two situations correspond to working conditionally on the chosen k or unconditionally with respect to k in a logic of multimodel inference. The latter estimate is useful when, as in this case, there is some uncertainty related to the number

TABLE 8.2
Acceptance rates for the MH, the split/combine, and the birth/death moves for $\sigma_\alpha^2 = \sigma_\psi^2 = 5, 10$

	$\sigma_\alpha^2 = \sigma_\psi^2 = 5$	$\sigma_\alpha^2 = \sigma_\psi^2 = 10$
	% Accepted	% Accepted
MH with fixed k		
Initial probabilities	19.21	18.64
Transition probabilities	18.18	16.17
$\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$, ψ_3	16.13	17.71
$\boldsymbol{\alpha}_u$	25.34	29.37
Birth	0.32	0.35
Death	0.33	0.36
Split	0.12	0.11
Combine	0.12	0.10

of latent states. In both tables, we also denote by two asterisks the 95% credible intervals not containing zero. Note that some covariates have been standardized in order to avoid numerical problems. On the basis of the estimates of the parameters for the covariates, we can see that the results are very similar either if we take the average of the draws limited to the value of k of interest or if we take the overall average.

On the basis of these estimation results, we conclude that age has an effect on fertility but not on employment. In this regard, we need to consider that the women in the sample were aged between 18 and 47, which is a limited range of years if we want to study the effect of aging on the probability of having a job. Education has a significant effect on both fertility and employment, whereas income of the husband affects only the logit of employment. Moreover, the number of children aged between 1 and 5 years has an effect on the employment, whereas the number of children aged between 3 and 13 years affects the fertility. The log-odds ratio between the two response variables is negative and is significant only on the basis of the 90% credible interval, rather than on the basis of the more standard 95% credible interval. This result can be interpreted as a negative contemporary association between the two response variables.

In Tables 8.5 and 8.6 we also show the estimates of the intercepts α_{1u} and α_{2u} (one for the marginal logit of fertility and the other for that of employment) corresponding to each latent state u, together with the estimated initial probabilities and transition probability matrix.

Though the number of states selected is different, both tables lead

TABLE 8.3
Posterior estimates of the model parameters $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \psi_3)$ affecting the marginal logits for fertility and employment and the log-odds ratio, computed as conditional averages of the draws given k

	Effect	$\sigma^2_\alpha = \sigma^2_\psi = 5$	$\sigma^2_\alpha = \sigma^2_\psi = 10$
		$k = 4$	$k = 5$
Logit fertility	race	-0.082	-0.082
	age†	1.434	1.680
	age^{2†}	-2.473^{**}	-2.736^{**}
	education†	0.369^{**}	0.363^{**}
	child 1-2	-0.019	-0.010
	child 3-5	-0.375^{**}	-0.381^{**}
	child 6-13	-0.673^{**}	-0.683^{**}
	child 14-	-0.454	-0.467
	income†	0.051	0.057
Logit employment	race	0.089	0.205
	age†	-0.780	-2.016
	age^{2†}	0.367	1.433
	education†	1.093^{**}	1.510^{**}
	child 1-2	-0.916^{**}	-1.147^{**}
	child 3-5	-0.706^{**}	-0.894^{**}
	child 6-13	-0.259	-0.377
	child 14-	0.363	0.354
	income†	-0.537^{**}	-0.696^{**}
Log-odds ratio	intercept	-1.165	-2.622

Note: † standardized covariate
** 95% credibility interval does not contain 0

to the same conclusions. The latent states are in increasing order on the basis of the marginal logit of employment; the latent states may therefore be interpreted as different levels of propensity of giving birth to a child and of getting a job position. For example, the first latent state corresponds to subjects with the highest propensity to fertility and the lowest propensity to employment. Moreover, it is interesting to observe that the transition matrix has an almost diagonal structure, with a large percentage of subjects that remain in the same latent state.

We finally outline the results of the application of the continuous BD algorithm (see Section 8.4.1) to the PSID dataset. Even if this algorithm may be considered less standard than the RJ algorithm, it represents an alternative and efficient sampling strategy that allows us to consider more complex models. In particular, through this algorithm we were able to deal, on the entire dataset of $n = 1446$ women, with the model

TABLE 8.4
Posterior estimates of the model parameters ($\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$, ψ_3) affecting the marginal logits for fertility and employment and the log-odds ratio, computed as overall averages of the draws

	Effect	$\sigma^2_\alpha = \sigma^2_\psi = 5$	$\sigma^2_\alpha = \sigma^2_\psi = 10$
		$k = 4$	$k = 5$
Logit fertility	race	-0.085	-0.080
	age†	1.339	1.725
	age^{2†}	-2.367^{**}	-2.793^{**}
	education†	0.364^{**}	0.367^{**}
	child 1-2	-0.021	-0.013
	child 3-5	-0.378^{**}	-0.373^{**}
	child 6-13	-0.679^{**}	-0.680^{**}
	child 14-	-0.471	-0.449
	income†	0.053	0.051
Logit employment	race	0.079	0.127
	age†	-0.849	-1.182
	age^{2†}	0.468	0.645
	education†	1.145^{**}	1.395^{**}
	child 1-2	-0.870^{**}	-1.151^{**}
	child 3-5	-0.704^{**}	-0.897^{**}
	child 6-13	-0.255	-0.384
	child 14-	0.352	0.382
	income†	-0.543^{**}	-0.669^{**}
Log-odds ratio	intercept	-1.594	-2.118

Note: ** 95% credibility interval does not contain 0
† standardized covariate

in a version that includes, among covariates, a time dummy for each year and the lagged response variables.

The resulting posterior probabilities of the number of latent states is in this case concentrated on $k = 3$, a result which is consistent with that found in Section 5.10.2. This may be due to the fact that we are using a larger sample and that we are removing an important part of the unobserved heterogeneity by modeling serial correlation through the introduction of the lagged responses among the covariates. The posterior estimates of the model parameters are instead quite similar to those obtained through the RJ algorithm applied to the restricted sample of $n = 482$ women. Then, we conclude that both algorithms are able to adequately approximate the posterior distributions, even if we have to acknowledge the higher efficiency of the BD algorithm. On the other hand, the parallel sampling algorithm, outlined in Section 8.4.2, is sim-

TABLE 8.5
Estimates of the intercepts for each latent state (α_{1u}, α_{2u}), of the initial probabilities (π_u), and of the transition probabilities ($\pi_{u|\bar{u}}$) for the prior hyperparameters $\sigma_\alpha^2 = \sigma_\psi^2 = 5$, with $k = 4$ latent states

	Support points		Initial				
u	Fertility	Empl.	prob.	Transition probabilities			
1	−1.796	−4.937	0.092	0.734	0.065	0.055	0.146
2	−1.936	−3.718	0.102	0.072	0.643	0.067	0.219
3	−2.648	−0.002	0.228	0.071	0.072	0.754	0.103
4	−2.609	5.980	0.578	0.021	0.027	0.009	0.944

TABLE 8.6
Estimates of the intercepts for each latent state (α_{1u}, α_{2u}), of the initial probabilities (π_u), and of the transition probabilities ($\pi_{u|\bar{u}}$) for the prior hyperparameters $\sigma_\alpha^2 = \sigma_\psi^2 = 10$, with $k = 5$ latent states

	Support points		Initial					
u	Fertility	Empl.	prob.	Transition probabilities				
1	−1.902	−5.461	0.084	0.553	0.063	0.090	0.060	0.234
2	−1.965	−5.117	0.103	0.043	0.754	0.050	0.052	0.101
3	−3.180	0.174	0.194	0.048	0.052	0.684	0.163	0.053
4	−2.117	3.542	0.090	0.187	0.081	0.059	0.555	0.117
5	−2.634	7.786	0.530	0.022	0.014	0.005	0.021	0.939

pler to implement for the present model, but requires running separate MCMC samplers for all values of k of interest and this may be considered a limitation.

A

Software

A.1 Introduction

We make available an R package that may be used to fit some of the models described in this book. This package is called LMest[1] and its use is described in the following section.

Other packages may be used in order to perform some of the analyses described in this book or similar analyses. In particular, we consider the following R packages[2]:

- lcmm: latent class mixed models;
- poLCA: latent class analysis for polytomous outcome variables;
- msm: multi-state Markov and hidden Markov models in continuous time;
- HiddenMarkov: discrete time hidden Markov models and related models.

Finally, a non-free software which can be used to fit certain latent class and latent Markov models is latent GOLD[3]; certain MATLAB packages are also available for this aim.

A.2 Package LMest

The main function of this package is est_lm_basic, which estimates, by the Expectation-Maximization algorithm described in Section 3.5.1, the

[1] the package may be directly installed in R through the compressed file http://www.stat.unipg.it/bartolucci/LMest_1.0.tar.gz

[2] freely downloadable from the website: http://www.R-project.org

[3] see http://www.statisticalinnovations.com/products/latentgold.html

basic LM model and some constrained versions of this model. In input, this function accepts the following arguments:

- `S`: array, of dimension $n \times T \times r$, of the distinct response configurations;
- `yv`: vector of frequencies in the sample of each response configuration;
- `k`: number of latent states;
- `start`: type of starting values (0 for deterministic, 1 for random); 0 is the default argument;
- `mod`: model on the transition probabilities (0 for time heterogeneous, 1 for time homogeneous, $2, \ldots, T$ for partial homogeneity of order `mod`); 0 is the default argument;
- `tol`: tolerance level for convergence; 10^{-6} is the default value.

In output, the function returns the following arguments:

- `lk`: maximum log-likelihood;
- `piv`: estimate of initial probability vector;
- `Pi`: estimate of transition probability matrices;
- `Psi`: estimate of conditional response probabilities;
- `np`: number of free parameters;
- `aic`: value of AIC for model selection;
- `bic`: value of BIC for model selection;
- `lkv`: log-likelihood trace at every EM step;
- `V`: array containing the posterior distribution of the latent states for each response configuration and time occasion;
- `Ul`: matrix containing the predicted sequences of latent states by the local decoding method.

In the following, we illustrate the main input arguments of the function `est_lm_basic` by referring to the example about the marijuana consumption dataset that is illustrated in Section 1.4.1; these data are available in the `LMest` package.

The main arguments are `S`, `yv`, and `k`. The last argument is simply equal to the number of latent states, such as 3 for these data. Moreover, in the case of univariate data, `S` is a matrix having a number of rows equal to the number of distinct response configurations and a number of columns equal to T, that is, the number of occasions of observation; `yv` is the corresponding vector of frequencies in the sample. For the marijuana consumption dataset, for instance, we have (see the help accompanying the package for the `R` commands required to load these data):

```
> S
      V1 V2 V3 V4 V5
 [1,]  0  0  0  0  0
 [2,]  0  0  0  0  1
 [3,]  0  0  0  0  2
 [4,]  0  0  0  1  0
 [5,]  0  0  0  1  1
 [6,]  0  0  0  1  2
 [7,]  0  0  0  2  0
 [8,]  0  0  0  2  1
 [9,]  0  0  0  2  2
[10,]  0  0  1  0  0
 ...................
[41,]  1  1  1  1  1
[42,]  1  1  2  2  2
[43,]  1  2  1  0  0
[44,]  1  2  1  2  2
[45,]  1  2  2  2  1
[46,]  1  2  2  2  2
[47,]  2  0  0  0  0
[48,]  2  1  2  2  2
[49,]  2  2  2  1  2
[50,]  2  2  2  2  0
[51,]  2  2  2  2  2
```

```
> yv
 [1] 111  18   7   6   6   1   2   1   4   2   ...
[35] ...      1   1   1   1   3   1   1   1   1   1
```

This means that, in this dataset, there are 51 distinct response configurations. For instance, the first configuration corresponds to response 0 provided at all occasions; this response configuration was provided by 111 subjects in the sample. The second configuration corresponds to all responses equal to 0 apart from the last one which is equal to 1; for this configuration, the frequency in the sample is equal to 18.

In order to run the estimation function for the data reported above, using 3 latent states and time-homogeneous transition matrices, we use the following command in `R`:

```
> out = est_lm_basic(S,yv,3,mod=1)
```

and to obtain the following output:

```
model type =       1
starting values =  0
n.latent states =  3
tolerance level =  1e-06
 |-------------|-------------|-------------|-------------|
 |     step    |     lk      |    lk-lko   | discrepancy |
 |-------------|-------------|-------------|-------------|
 |           0 |    -825.611 |
 |          10 |      -669.9 |     1.99854 |   0.0265126 |
 |          20 |    -661.029 |    0.425886 |   0.0111137 |
 |          30 |    -659.057 |   0.0833311 |  0.00484088 |
 |          40 |    -658.698 |   0.0149382 |  0.00188596 |
 |          50 |    -658.626 |  0.00362283 | 0.000891454 |
 |          60 |    -658.606 |  0.00116616 | 0.000467175 |
 |          70 |    -658.599 | 0.000455424 | 0.000263994 |
 |          80 |    -658.596 |  0.00020406 | 0.000157066 |
 |          90 |    -658.595 | 0.000102188 | 9.67821e-05 |
 |         100 |    -658.594 | 5.64752e-05 | 6.10552e-05 |
 |         110 |    -658.593 | 3.40552e-05 |  3.9111e-05 |
 |         120 |    -658.593 | 2.20731e-05 | 2.52908e-05 |
 |         130 |    -658.593 | 1.51138e-05 |  1.6438e-05 |
 |         140 |    -658.593 |  1.0756e-05 | 1.07044e-05 |
 |         150 |    -658.593 | 7.85353e-06 | 6.96631e-06 |
 |         160 |    -658.593 |  5.8291e-06 | 4.52086e-06 |
 |         170 |    -658.593 | 4.37114e-06 | 2.91958e-06 |
 |         180 |    -658.593 | 3.29876e-06 | 1.87219e-06 |
 |         190 |    -658.592 | 2.49932e-06 | 1.28873e-06 |
 |         200 |    -658.592 | 1.89829e-06 | 9.16588e-07 |
 |         210 |    -658.592 | 1.44404e-06 | 6.54581e-07 |
 |         220 |    -658.592 | 1.09959e-06 | 4.69284e-07 |
 |         224 |    -658.592 | 9.86231e-07 | 4.37347e-07 |
 |-------------|-------------|-------------|-------------|
```

In each line of the above output we have the number of the iteration of the EM algorithm used to maximize the model log-likelihood, the log-likelihood level at the end of the iteration, the difference with respect to the log-likelihood at the end of the previous iteration, and the discrepancy between the corresponding parameter vectors.

The function returns a list of objects, that in the specific case is contained in `out`. In fact, we have

```
> names(out)
[1] "lk"  "piv" "Pi"  "Psi" "np"  "aic" "bic" "lkv" "V"
```

For the marijuana consumption dataset, in particular, we have

```
> out
$lk
[1] -658.5924

$piv
[1] 0.91216458 0.07115947 0.01667595

$Pi
, , 1

     [,1] [,2] [,3]
[1,]    0    0    0
[2,]    0    0    0
[3,]    0    0    0

, , 2

             [,1]      [,2]      [,3]
[1,] 8.417068e-01 0.1408092 0.0174840
[2,] 8.020766e-02 0.6697948 0.2499975
[3,] 2.694993e-44 0.1319117 0.8680883

.....................................

, , 5

             [,1]      [,2]      [,3]
[1,] 8.417068e-01 0.1408092 0.0174840
[2,] 8.020766e-02 0.6697948 0.2499975
[3,] 2.694993e-44 0.1319117 0.8680883
```

```
$Psi
, , 1

            [,1]       [,2]         [,3]
[1,] 0.988820196 0.28922288 1.018329e-05
[2,] 0.007158569 0.67903172 5.244646e-02
[3,] 0.004021234 0.03174539 9.475434e-01


$np
[1] 14

$aic
[1] 1345.185

$bic
[1] 1393.738

$lkv
  [1] -712.5494 -693.1446 -686.9130 -683.9249 -681.4998 ...
  .........................................................
[222] -658.5924 -658.5924 -658.5924

$V
, , 1

              [,1]         [,2]         [,3]
             [,1]        [,2]         [,3]
 [1,] 0.997147368 0.002852631 1.166185e-09
 [2,] 0.996292040 0.003707957 2.626665e-09
 [3,] 0.995551455 0.004448541 3.891221e-09
 [4,] 0.993590242 0.006409751 7.240012e-09
 [5,] 0.993055469 0.006944523 8.153143e-09
.........................................
[46,] 0.009316909 0.933445079 5.723801e-02
[47,] 0.906406393 0.069986504 2.360710e-02
[48,] 0.068065495 0.211048655 7.208858e-01
[49,] 0.004893540 0.040433396 9.546731e-01
[50,] 0.004838734 0.040308593 9.548527e-01
[51,] 0.004832637 0.040295511 9.548719e-01

.........................................
```

```
, , 5

               [,1]        [,2]         [,3]
 [1,] 9.405806e-01 0.059419052 3.900707e-07
 [2,] 4.592593e-02 0.940886377 1.318769e-02
 [3,] 8.374849e-02 0.142794609 7.734569e-01
 [4,] 3.797760e-01 0.620215313 8.694530e-06
 [5,] 1.829920e-03 0.969162322 2.900776e-02
.........................................
[46,] 2.063053e-06 0.005192057 9.948059e-01
[47,] 9.375203e-01 0.062479307 4.333337e-07
[48,] 2.531663e-06 0.005216338 9.947811e-01
[49,] 4.828326e-04 0.034925286 9.645919e-01
[50,] 1.050383e-02 0.989266402 2.297716e-04
[51,] 2.042106e-06 0.005190976 9.948070e-01
```

All the above objects have a rather simple interpretation. Note, in particular, that `out$Pi[,,t]`, with `t` between 2 and the number of times occasions, corresponds to a specific transition probability matrix. Moreover, `out$V[,,t]` contains the corresponding posterior distribution of the latent states for every response configuration. For instance, we have

```
> out$V[1,,2]
[1] 9.928663e-01 7.133706e-03 2.174017e-09
```

which corresponds to the posterior distribution of the latent states at the second occasion of interview, given the first response configuration. By the method of local decoding, see Section 7.5.1, subjects with this response configuration are assigned to the first latent state at the second occasion, as this state corresponds to the largest element in the above vector.

In order to fit the same model as above with time-heterogenous transition matrices, we have to include the argument `mod=0` (or simply remove this argument), obtaining in this way the same estimates illustrated in Section 3.7.1. It is also possible to use the constraint of partial homogeneity, as described in Section 4.3. Similarly, to use random instead of deterministic starting values for the EM algorithm (see Section 3.5.1.3), we have to include the input argument `start=1`.

Another important function that is included in the `LMest` package is `bootstrap_lm_basic`, which performs a parametric bootstrap to obtain the standard errors for the parameter estimates (see Section 7.4.2). This function accepts the following input arguments:

- `piv`: initial probability vector;

- `Pi`: transition probability matrices;
- `Psi`: matrix of conditional response probabilities;
- `n`: sample size;
- `B`: number of bootstrap samples (100 by default);
- `start`: see the same input for `est_lm_basic`;
- `mod`: see the same input for `est_lm_basic`;
- `tol`: see the same input for `est_lm_basic`.

Moreover, the function returns the following output arguments:

- `mPsi`: average of bootstrap estimates of the conditional response matrix;
- `mpiv`: average of bootstrap estimates of the initial probability vector;
- `mPi`: average of bootstrap estimates of the transition probability matrices;
- `sePsi`: standard errors for the conditional response matrix;
- `sepiv`: standard errors for the initial probability vector;
- `sePi`: standard errors for the transition probability matrices.

Again with reference to the marijuana consumption dataset and using the output provided by the function `est_lm_basic`, we can obtain the standard errors for the parameter estimates, based on 100 bootstrap samples, by the following `R` command:

```
> out_boot = bootstrap_lm_basic(out$piv,out$Pi,out$Psi,n,mod=1)
```

where `n` is set equal to the sample size, that is, 237.

The main objects contained in the list returned by the function are the following:

```
> names(out)
[1] "mPsi"  "mpiv"  "mPi"   "sePsi" "sepiv" "sePi"
```

In particular, for the example at issue we have the following output argument, the interpretation of which is obvious:

```
> out_boot
$mPsi
, , 1

           [,1]       [,2]       [,3]
[1,] 0.98786388 0.27394409 0.01310875
[2,] 0.00787959 0.69347075 0.07192772
[3,] 0.00425653 0.03258516 0.91496353


$mpiv
[1] 0.91166790 0.07152271 0.01680939

$mPi
, , 1

     [,1] [,2] [,3]
[1,]    0    0    0
[2,]    0    0    0
[3,]    0    0    0

, , 2

           [,1]      [,2]       [,3]
[1,] 0.84074650 0.1396729 0.01958056
[2,] 0.07568413 0.6758765 0.24843935
[3,] 0.00905775 0.0773157 0.91362655

....................................

, , 5

           [,1]      [,2]       [,3]
[1,] 0.84074650 0.1396729 0.01958056
[2,] 0.07568413 0.6758765 0.24843935
[3,] 0.00905775 0.0773157 0.91362655
```

```
$sePsi
, , 1

            [,1]       [,2]       [,3]
[1,] 0.009572456 0.08124714 0.01898721
[2,] 0.008512781 0.08210117 0.05734771
[3,] 0.004087386 0.03239681 0.06140618


$sepiv
[1] 0.023965373 0.024711894 0.009161026

$sePi
, , 1

     [,1] [,2] [,3]
[1,]    0    0    0
[2,]    0    0    0
[3,]    0    0    0

, , 2

           [,1]       [,2]       [,3]
[1,] 0.01740978 0.02044865 0.01065115
[2,] 0.05090068 0.06909400 0.04616496
[3,] 0.02089930 0.06511651 0.06571420

.....................................

, , 5

           [,1]       [,2]       [,3]
[1,] 0.01740978 0.02044865 0.01065115
[2,] 0.05090068 0.06909400 0.04616496
[3,] 0.02089930 0.06511651 0.06571420
```

In particular, `out_boot$sePsi` contains the standard errors for the estimates in `out$Psi`, `out_boot$sepiv` contains the standard errors for those in `out$piv`, and `out_boot$sePi` contains the standard errors for those in `out$Pi`. See also Section 7.4.2 for the results obtained under the model with time-heterogenous transition matrices.

An important point concerns how to perform local and global decoding; see Sections 7.5.1 and 7.5.2. The first task is already performed by the estimation function `est_lm_basic` and the predicted sequences are in the matrix `Ul`. Then, after estimation, for the same data about marijuana consumption we have

```
> out$Ul
      [,1] [,2] [,3] [,4] [,5]
 [1,]    1    1    1    1    1
 [2,]    1    1    1    1    2
 [3,]    1    1    1    1    3
 [4,]    1    1    1    2    2
 [5,]    1    1    1    2    2
 [6,]    1    1    1    2    3
 [7,]    1    1    1    1    1
 [8,]    1    1    1    3    2
 [9,]    1    1    1    3    3
[10,]    1    1    2    1    1
..............................
[41,]    2    2    2    2    2
[42,]    2    2    3    3    3
[43,]    2    3    2    1    1
[44,]    2    3    3    3    3
[45,]    2    3    3    3    2
[46,]    2    3    3    3    3
[47,]    1    1    1    1    1
[48,]    3    2    3    3    3
[49,]    3    3    3    3    3
[50,]    3    3    3    3    2
[51,]    3    3    3    3    3
```

Note that the matrix `Ul` has the same dimension as `S`. Every row of `Ul` contains the sequence of latent states predicted for the corresponding row of the second matrix. So, for instance, the first row `Ul` contains the sequence of latent states corresponding to the response configuration made of all zeros.

In order to perform global decoding, in the package `LMest` there is the specific function `viterbi`. The input arguments of the function are

- `S`: array, of dimension $n \times T \times r$, of the distinct response configurations;
- `piv`: initial probability vector;
- `Pi`: probability transition matrices;
- `Psi`: matrix of conditional response probabilities.

The following output argument is returned:

- `U`: matrix containing the predicted sequences of latent states by the global decoding method.

For instance, for the data analyzed above, we can perform global decoding by the command

```
out2 = viterbi(S,out$piv,out$Pi,out$Psi)
```

Then, we obtain the following output:

```
> out2$U
      [,1] [,2] [,3] [,4] [,5]
 [1,]    1    1    1    1    1
 [2,]    1    1    1    1    2
 [3,]    1    1    1    1    3
 [4,]    1    1    1    2    2
 [5,]    1    1    1    2    2
 [6,]    1    1    1    2    3
 [7,]    1    1    1    1    1
 [8,]    1    1    1    2    2
 [9,]    1    1    1    3    3
[10,]    1    1    2    1    1
.............................
[41,]    2    2    2    2    2
[42,]    2    2    3    3    3
[43,]    2    3    2    1    1
[44,]    2    3    3    3    3
[45,]    2    3    3    3    2
[46,]    2    3    3    3    3
[47,]    1    1    1    1    1
[48,]    3    3    3    3    3
[49,]    3    3    3    3    3
[50,]    3    3    3    3    2
[51,]    3    3    3    3    3
```

The returned matrix is organized exactly as the matrix `Ul` containing the sequences predicted by the local decoding method.

Finally, we clarify that the package `LMest` contains another dataset for performing examples of application of the multivariate LM model. How to perform these examples is illustrated in the package help. Moreover, the package contains other functions to manage longitudinal data. The authors will continue to include, in the same package, other `R` functions to estimate more sophisticated versions of the LM models, even including covariates.

List of Main Symbols

Below we present a list of the principal symbols used in the main part of the book, from Chapter 3 through Chapter 6; the symbols are listed in the same order as they are introduced in these chapters and, when useful, we distinguish the univariate case (one response variable for each time occasion) from the multivariate case (more response variables for each time occasion). We recall that, throughout the book, we denote by $f_A(a)$ the probability mass (or density) function of the distribution of the random variable A and by $f_{A|B}(a|b)$ that of the conditional distribution of A given the random variable B. A similar notation applies for random vectors. Moreover, vectors are usually indicated in bold (by a capital letter in the case of a random vector and by a small letter for one of its realizations), whereas matrices are indicated by capital bold letters. Finally, parameters are indicated by greek letters and the symbol "hat" generally indicates an estimate of a certain quantity.

Symbol	Description
T	number of time occasions
$Y^{(t)}$	response variable at occasion t (univ. case)
c	number of categories of each response variable (univ. case)
$\tilde{\boldsymbol{Y}}$	vector of all response variables (univ./multiv. cases)
$U^{(t)}$	latent variable at occasion t
k	number of latent states
$\boldsymbol{U}$	latent process $(U^{(1)}, \dots, U^{(T)})$
$\phi_{y\|u}$, $\phi_{y\|u}^{(t)}$	conditional response probability $f_{Y^{(t)}\|U^{(t)}}(y\|u)$ (univ. case)
π_u	initial probability $f_{U^{(1)}}(u)$
$\pi_{u\|\bar{u}}^{(t)}$	transition probability $f_{U^{(t)}\|U^{(t-1)}}(u\|\bar{u})$
#par	number of free parameters
$q^{(t)}(u, \tilde{\boldsymbol{y}})$	element $f_{U^{(t)},Y^{(1)},\dots,Y^{(t)}}(u, y^{(1)}, \dots, y^{(t)})$ or $f_{U^{(t)},\boldsymbol{Y}^{(1)},\dots,\boldsymbol{Y}^{(t)}}(u, \boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(t)})$ of the forward recursion
$q^{(t)}(u)$	element $f_{U^{(t)}}(u)$ of the forward recursion
r	number of responses for each time occasion (multiv. case)
$Y_j^{(t)}$	j-th response variable at occasion t (multiv. case)
c_j	number of categories of the j-th response variable (multiv. case)

Symbol	Description
$\boldsymbol{Y}^{(t)}$	vector of response variables at occasion t (multiv. case)
$\phi_{jy\|u}$, $\phi_{jy\|u}^{(t)}$	conditional response probabilities $f_{Y_j^{(t)}\|U^{(t)}}(y\|u)$ (multiv. case)
$\phi_{\boldsymbol{y}\|u}$, $\phi_{\boldsymbol{y}\|u}^{(t)}$	conditional response probability $f_{\boldsymbol{Y}^{(t)}\|U^{(t)}}(\boldsymbol{y}\|u)$ (multiv. case)
n	sample size
$y_i^{(t)}$	response provided by sample unit i at occasion t (univ. case)
$y_{ij}^{(t)}$	response of type j provided by sample unit i at occasion t (multiv. case)
$\boldsymbol{y}_i^{(t)}$	vector of responses provided by sample unit i at occasion t (multiv. case)
$\tilde{\boldsymbol{y}}_i$	observed response configuration for sample unit i (univ./multiv. cases)
$\boldsymbol{\theta}$	vector of model parameters
$\ell(\boldsymbol{\theta})$	model log-likelihood
$n_{\tilde{\boldsymbol{y}}}$	frequency of the response configuration $\tilde{\boldsymbol{y}}$
$\ell^*(\boldsymbol{\theta})$	complete data log-likelihood
$\bar{q}^{(t)}(\bar{u},\tilde{\boldsymbol{y}})$	element $f_{Y^{(t+1)},\ldots,Y^{(T)}\|U^{(t)}}(y^{(t+1)},\ldots,y^{(T)}\|\bar{u})$ or $f_{\boldsymbol{Y}^{(t+1)},\ldots,\boldsymbol{Y}^{(T)}\|U^{(t)}}(\boldsymbol{y}^{(t+1)},\ldots,\boldsymbol{y}^{(T)}\|\bar{u})$ of the backward recursion
AIC	index used in the Akaike information criterion
BIC	index used in the Bayesian information criterion
$\hat{\ell}$	maximum log-likelihood of the model of interest
S	index which measures the quality of the classification
$\mathbf{0}_h$	column vector of h zeros
$\mathbf{1}_h$	column vector of h ones
$\boldsymbol{O}_{hj}$	$h \times j$ matrix of zeros
$\boldsymbol{I}_h$	identity matrix of size h
$\boldsymbol{d}_{hj}$	column vector of j zeros with the h-th element equal to one
$\otimes$	Kronecker product
$\boldsymbol{\phi}_u^{(t)}$	vector with elements $\phi_{yu}^{(t)}$ (univ. case)
$\boldsymbol{\eta}_u^{(t)}$	vector of logits for $\boldsymbol{\phi}_u^{(t)}$ (univ. case)
$\boldsymbol{g}(\cdot)$	link function between $\boldsymbol{\phi}_u^{(t)}$ and $\boldsymbol{\eta}_u^{(t)}$ (univ. case)
$\boldsymbol{W}_u^{(t)}$	design matrix for the model on the conditional response probabilities (univ. case)
$\boldsymbol{\beta}$	vector of parameters of the model on the conditional response probabilities (univ./multiv. case)
$\eta_{y\|u}^{(t)}$	single element of $\boldsymbol{\eta}_u^{(t)}$ when $c > 1$ (univ. case)
$\boldsymbol{\phi}_{j\|u}^{(t)}$	vector with elements $\phi_{jy\|u}^{(t)}$ (multiv. case)
$\boldsymbol{\eta}_{j\|u}^{(t)}$	vector of logits for $\boldsymbol{\phi}_{j\|u}^{(t)}$ (multiv. case)

Symbol	Description
$\boldsymbol{g}_j(\cdot)$	link function between $\boldsymbol{\phi}^{(t)}_{j\|u}$ and $\boldsymbol{\eta}^{(t)}_{j\|u}$ (multiv. case)
$\boldsymbol{W}^{(t)}_{j\|u}$	design matrix for the model on the conditional response probabilities (multiv. case)
T^*	time occasion separating two periods with different transition probabilities (constraint of partial homogeneity)
$\pi^{*(h)}_{u\|\bar{u}}$	transition probabilities (constraint of partial homogeneity, $h = 1, 2$)
$\boldsymbol{\rho}^{(t)}_{\bar{u}}$	vector of the transition probabilities $\pi^{(t)}_{u\|\bar{u}}$ for $u \neq \bar{u}$
$\boldsymbol{Z}^{(t)}_{\bar{u}}$	design matrix for the (linear or GLM) model on the transition probabilities
$\boldsymbol{\delta}$	vector of parameters of the (linear or GLM) model on the transition probabilities
$\boldsymbol{\pi}^{(t)}_{\bar{u}}$	vector with elements $\pi^{(t)}_{u\|\bar{u}}$
$\boldsymbol{\lambda}^{(t)}_{\bar{u}}$	vector of logits for $\boldsymbol{\pi}^{(t)}_{u}$
LR	likelihood ratio statistic
$\boldsymbol{T}_h$	$h \times h$ lower triangular matrix of ones
$\boldsymbol{X}^{(t)}$	vector of covariates at time occasion t
$\tilde{\boldsymbol{X}}$	vector of all individual covariates
$\phi^{(t)}_{y\|u\boldsymbol{x}}$	conditional response probability $f_{Y^{(t)}\|U^{(t)},\boldsymbol{X}^{(t)}}(y\|u,\boldsymbol{x})$ (univ. case)
$\phi^{(t)}_{jy\|u\boldsymbol{x}}$	conditional response probability $f_{Y_j^{(t)}\|U^{(t)},\boldsymbol{X}^{(t)}}(y\|u,\tilde{\boldsymbol{x}})$ (multiv. case)
$\pi_{u\|\boldsymbol{x}}$	initial probability $f_{U^{(t)}\|\boldsymbol{X}^{(t)}}(u\|\boldsymbol{x})$
$\pi^{(t)}_{u\|\bar{u}\boldsymbol{x}}$	transition probability $f_{U^{(t)}\|U^{(t-1)},\boldsymbol{X}^{(t)}}(u\|\bar{u},\boldsymbol{x})$
$\phi^{(t)}_{\boldsymbol{y}\|u\boldsymbol{x}}$	conditional response probability $f_{\boldsymbol{Y}^{(t)}\|U^{(t)},\boldsymbol{X}^{(t)}}(\boldsymbol{y}\|u,\boldsymbol{x})$
$\boldsymbol{\phi}^{(t)}_{u\boldsymbol{x}}$	vector with elements $\phi^{(t)}_{Jy\|u\boldsymbol{x}}$ (univ. case)
$\boldsymbol{\eta}^{(t)}_{u\boldsymbol{x}}$	vector of logits for $\boldsymbol{\phi}^{(t)}_{u\boldsymbol{x}}$ (univ. case)
$\boldsymbol{W}^{(t)}_{u\boldsymbol{x}}$	design matrix for the model on the conditional response probabilities (univ. case)
$\boldsymbol{\phi}^{(t)}_{j\|u\boldsymbol{x}}$	vector with elements $\phi^{(t)}_{jy\|u\boldsymbol{x}}$ (multiv. case)
$\boldsymbol{\eta}^{(t)}_{j\|u\boldsymbol{x}}$	vector of logits for $\boldsymbol{\phi}^{(t)}_{j\|u\boldsymbol{x}}$ (multiv. case)
$\boldsymbol{W}^{(t)}_{j\|u\boldsymbol{x}}$	design matrix for the model on the conditional response probabilities (multiv. case)
$\boldsymbol{\pi}_{\boldsymbol{x}}$	initial probability vector with elements $\pi_{u\|\boldsymbol{x}}$
$\boldsymbol{\lambda}_{\boldsymbol{x}}$	vector of logits for $\boldsymbol{\pi}_{\boldsymbol{x}}$
$\boldsymbol{Z}_{\boldsymbol{x}}$	design matrix for the model on the initial probabilities
$\boldsymbol{\gamma}$	vector of parameters for the model on the initial probabilities
$\boldsymbol{\pi}_{\bar{u}\boldsymbol{x}}$	vector of transition probabilities $\pi_{u\|\bar{u}\boldsymbol{x}}$

Symbol	Description
$\boldsymbol{\lambda}^{(t)}_{\bar{u}\boldsymbol{x}}$	vector of logits for $\boldsymbol{\pi}_{\bar{u}\boldsymbol{x}}$
$\boldsymbol{Z}^{(t)}_{\bar{u}\boldsymbol{x}}$	design matrix of the model on the transition probabilities
$\boldsymbol{x}^{(t)}_{i}$	observed configuration of the covariates for sample unit i at occasion t
$\tilde{\boldsymbol{x}}_i$	observed covariates for sample unit i
$n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}}$	frequency of covariate configuration $\tilde{\boldsymbol{x}}$ and response configuration $\tilde{\boldsymbol{y}}$
$\boldsymbol{J}(\boldsymbol{\theta})$	observed information matrix
$\boldsymbol{s}(\boldsymbol{\theta})$	score vector
$\mathrm{se}(\hat{\boldsymbol{\theta}})$	standard errors
$\boldsymbol{X}^{*(t)}$	vector of covariates at time occasion t (serial dependence case)
$\tilde{\boldsymbol{X}}^{*}$	vector of all the covariates (serial dependence case)
$\phi^{(t)}_{y\|u\boldsymbol{x}^*}$	conditional response probability $f_{Y^{(t)}\|U,\boldsymbol{X}^{(t)},Y^{(t-1)}}(y\|u,\boldsymbol{x},\bar{y})$ (serial dependence case)
$\boldsymbol{W}_{u\boldsymbol{x}^*}$	design matrix for the model on the conditional response probabilities (serial dependence case)
$\boldsymbol{p}^{(t)}_{u\boldsymbol{x}}$	column vector with elements $\phi^{(t)}_{\boldsymbol{y}\|u\boldsymbol{x}}$ (contemporary dependence case)
$\boldsymbol{\eta}^{*(t)}_{u\boldsymbol{x}}$	vector of logits for $\boldsymbol{p}^{(t)}_{u\boldsymbol{x}}$ (contemporary dependence case)
$\boldsymbol{g}^{*}(\cdot)$	multivariate link function for $\boldsymbol{p}^{(t)}_{u\boldsymbol{x}}$ (contemporary dependence case)
k''	number of the latent states (2nd-order model)
$\phi''_{y\|u''}$	conditional response probability $f_{Y^{(t)}\|U^{(t)}}(y\|u'')$ (2nd-order model, univ.case)
$\phi''_{jy\|u''}$	conditional response probability $f_{Y^{(t)}\|U^{(t)}}(y\|u'')$ (2nd-order model, multiv.case)
$\pi''_{u''}$	initial probability $f_{U^{(1)}}(u'')$ (2nd-order model)
$\pi''^{(2)}_{u''\|\bar{u}''}$	transition probability $f_{U^{(2)}\|U^{(1)}}(u''\|\bar{u}'')$ (2nd-order model)
$\pi''^{(t)}_{u''\|\bar{\bar{u}}''\bar{u}''}$	transition probability $f_{U^{(t)}\|U^{(t-2)},U^{(t-1)}}(u''\|\bar{\bar{u}}'',\bar{u}'')$ (2nd-order model)
U_1	random effect (random-effects model)
k_1	number of support points of the random-effects distribution (random-effects model)
$U_2^{(t)}$	latent variable at occasion t (random-effects model)
k_2	number of latent states (random-effects model)
$\boldsymbol{U}_2$	latent process $(U_2^{(1)},\ldots,U_2^{(T)})$ (random-effects model)
ω_{u_1}	mass probability for the random-effects distribution (random-effects model)
$\phi^{(t)}_{y\|u_1u_2\boldsymbol{x}}$	conditional response probability $f_{Y^{(t)}\|U_1,U_2^{(t)},\boldsymbol{X}^{(t)}}(y\|u_1,u_2,\boldsymbol{x})$ (random-effects model)

Symbol	Description
$\pi_{u_2\|u_1\boldsymbol{x}}$	initial probability $f_{U_2^{(1)}\|U_1,\boldsymbol{X}^{(t)}}(u_2\|u_1,\boldsymbol{x})$ (random-effects model)
$\pi^{(t)}_{u_2\|u_1\bar{u}_2\boldsymbol{x}}$	transition probability $f_{U_2^{(t)}\|U_1,U_2^{(t-1)},\boldsymbol{X}^{(t)}}(u_2\|u_1,\bar{u}_2,\boldsymbol{x})$ (random-effects model)
H	number of clusters (multilevel model)
n_h	number of units in cluster h (multilevel model)
$Y^{(t)}_{hi}$	response variable at occasion t for sample unit i in cluster h (multilevel model)
$\tilde{\boldsymbol{Y}}_{hi}$	vector of responses for subject i in cluster h (multilevel model)
$\tilde{\boldsymbol{Y}}_h$	vector of all responses for all subjects in cluster h (multilevel model)
$\tilde{\boldsymbol{X}}_{1h}$	vector of covariates for cluster h (multilevel model)
$\boldsymbol{X}^{(t)}_{2hi}$	vector of covariates for subject i in cluster h at time t (multilevel model)
$\tilde{\boldsymbol{X}}_{2hi}$	vector of all covariates for subject i in cluster h (multilevel model)
$\tilde{\boldsymbol{X}}_{2h}$	vector of all covariates for the subjects in cluster h (multilevel model)
U_{1h}	random effect for cluster h (multilevel model)
k_1	number of support points of the random-effects distribution (multilevel model)
$U^{(t)}_{2hi}$	latent variable at occasion t (multilevel model)
k_2	number of latent states (multilevel model)
$\boldsymbol{U}_{2hi}$	latent process $(U^{(1)}_{2hi},\ldots,U^{(T)}_{2hi})$ for subject i in cluster h (multilevel model)
$\phi^{(t)}_{y\|u_2}$	conditional response probability $f_{Y^{(t)}_{hi}\|U^{(t)}_{2hi}}(y\|u_2)$ (multilevel model)
$\omega_{u_1\|\boldsymbol{x}_1}$	mass probability $f_{U_{1h}\|\boldsymbol{X}_{1h}}(u_1\|\boldsymbol{x}_1)$ for the random-effects distribution (multilevel model)

Bibliography

A. Agresti (2002). *Categorical Data Analysis, 2nd Edition*. John Wiley & Sons, Hoboken, NJ.

H. Akaike (1973). Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov and F. Csáki, eds., *Second International Symposium of Information Theory*, 267–281. Akadémiai Kiado, Budapest.

R. M. Altman (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, **102**, 201–210.

R. M. Altman and A. J. Petkau (2005). Application of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine*, **24**, 2335–2344.

T. W. Anderson (1951). Probability models for analysing time changes in attitudes. In: P. F. Lazarsfelsd, ed., *The use of mathematical models in the measurement of the attitudes*. The RAND Research Memorandum No. 455.

T. W. Anderson (1954). Probability models for analysing time changes in attitudes. In: P. F. Lazarsfelsd, ed., *Mathematical Thinking in the Social Science*. The Free Press.

D. W. K. Andrews and M. Buchinsky (2000). A three-step method for choosing the number of bootstrap repetitions. *Econometrica*, **68**, 23–52.

T. Asparouhov and B. Muthén (2008). Multilevel mixture models. In: G. R. Hancock and K. M. Samuelson, eds., *Advances in Latent Variable Mixture Models*. Information Age Publishing, Charlotte, NC, pp. 27–51.

K. Auranen, E. Arjas, T. Leino, and A. K. Takala (2000). Transmission of pneumococcal carriage in families: a latent Markov process model for binary data. *Journal of the American Statistical Association*, **95**, 1044–1053.

A. Azzalini (1996). *Statistical Inference Based on the Likelihood*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Boca Raton, FL.

K. Bandeen-Roche, D. L. Miglioretti, S. L. Zeger, and P.J. Rathouz (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, **92**, 1375–1386.

D. J. Bartholomew, M. Knott, and I. Moustaki (2011). *Latent Variable Models and Factor Analysis: A Unified Approach, 3rd Edition*. Arnold, Chichester, UK.

F. Bartolucci (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society, series B*, **68**, 155–178.

F. Bartolucci (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, **72**, 141–157.

F. Bartolucci and J. Besag (2002). A recursive algorithm for Markov random fields. *Biometrika*, **89**, 724–730.

F. Bartolucci and A. Farcomeni (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, **104**, 816–831.

F. Bartolucci and A. Forcina (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika*, **70**, 31–43.

F. Bartolucci and A. Forcina (2006). A class of latent marginal models for capture-recapture data with continuous covariates. *Journal of the American Statistical Association*, **101**, 786–794.

F. Bartolucci, A. Forcina, and V. Dardanoni (2001). Positive quadrant dependence and marginal modelling in two-way tables with ordered margins. *Journal of the American Statistical Association*, **96**, 1497–1505.

F. Bartolucci, M. Lupparelli, and G. E. Montanari (2009). Latent Markov model for binary longitudinal data: an application to the performance evaluation of nursing homes. *Annals of Applied Statistics*, **3**, 611–636.

F. Bartolucci and F. Pennoni (2007). A class of latent Markov models for capture-recapture data allowing for time, heterogeneity and behavior effects. *Biometrics*, **63**, 568–578.

F. Bartolucci, F. Pennoni, and B. Francis (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society, Series A*, **170**, 151–132.

F. Bartolucci, F. Pennoni, and M. Lupparelli (2008). Likelihood inference for the latent Markov Rasch model. In: C. Huber, N. Limnios, M. Mesbah, and M. Nikulin, eds., *Mathematical Methods for Survival Analysis, Reliability and Quality of Life*, 239–254. Wiley, London.

F. Bartolucci, F. Pennoni, and G. Vittadini (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, **36**, 491–522.

F. Bartolucci, L. Scaccia, and A. Mira (2006). Efficient Bayes factor estimation from the reversible jump output. *Biometrika*, **93**, 41–52.

F. Bartolucci and I. L. Solis-Trapala (2010). Multidimensional latent Markov models in a developmental study of inhibitory control and attentional flexibility in early childhood. *Psychometrika*, **75**, 725–743.

L. E. Baum and J. A. Egon (1967). An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bullettin of the American Meteorological Society*, **73**, 360–363.

L. E. Baum and T. Petrie (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, **37**, 1554–1563.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.

A. Berchtold (2004). Optimization of mixture models: comparison of different strategies. *Computational Statistics*, **19**, 385–406.

J. M. Bernardo and A. F. M. Smith (1994). *Bayesian Theory*. Wiley, Chichester.

C. J. H. Bijleveld and A. Mooijaart (2003). Latent Markov modelling of recidivism data. *Statistica Neerlandica*, **57**, 305–320.

D. Böhning, E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**, 373–388.

K. A. Bollen and P. J. Curran (2006). *Latent Curve Models: A Structural Equation Perspective*. Wiley, Hoboken, NJ.

S. Boucheron and E. Gassiat (2005). An information-theoretic perspective on order estimation. In: O. Cappé, E. Moulines, and T. Rydén, ed., *Inference in Hidden Markov Models*, 565–601. Springer, New York.

A. S. Bryk and H. I. Weisberg (1976). Value-added analysis: a dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, **1**, 127–155.

K. P. Burnham and D. R. Anderson (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach.* 2nd Edition. Springer-Verlag, New York.

B. V. Bye and E. S. Schechter (1986). A latent Markov model approach to the estimation of response errors in multiwave panel data. *Journal of the American Statistical Association*, **81**, 375–380.

O. Cappé and E. Moulines (2005). Recursive computation of the score and observed information matrix in hidden Markov models. In: *13th IEEE Workshop on Statistical Signal Processing*, pp. 703–707.

O. Cappé, E. Moulines, and T. Rydén (2005). *Inference in Hidden Markov Models*. Springer, New York.

R. Carrasco (2001). Binary choice with binary endogenous regressors in panel data: estimating the effect of fertility on female labor participation. *Journal of Business and Economic Statistics*, **19**, 385–394.

G. Casella and L. R. Berger (2002). *Statistical Inference, 2nd Edition*. Duxbury Press, Pacific Grove, CA.

G. Celeux (1998). Bayesian inference for mixtures: the label-switching problem. In: R. Payne and P.J. Green, eds., *COMPSTAT 98-Proc. in Computational Statistics*, 227–232. Physica, Heidelberg.

G. Celeux and J. B. Durand (2008). Selecting hidden Markov chain states number with cross-validated likelihood. *Computational Statistics*, **23**, 541–564.

G. Celeux, M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, **95**, 957–970.

G. Celeux and G. Soromenho (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195–212.

M. R. Chernick (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers, 2nd Edition*. Wiley, Hoboken, NJ.

S. Chib (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, **75**, 79–97.

C. C. Clogg (1995). Latent class models. In: G. Arminger, C. C. Clogg, and M. E. Sobel, eds., *Handbook of Statistical Modelling for the Social and Behavioral Sciences*, 311–359. Plenum Press, New York.

I. B. Collings and T. Rydén (1998). A new maximum likelihood gradient algorithm for on-line hidden Markov model identification. In: *International Conference on Acoustics, Speech, and Signal Processing*, pp. 2261–2264.

L. M. Collins and S. T. Lanza (2010). *Latent Class and Latent Transition Analysis with Applications in the Social, Behavioural, and Health Sciences*. Wiley, Hoboken, NJ.

L. M. Collins and S. E. Wugalter (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, **27**, 131–157.

R. Colombi and A. Forcina (2001). Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, **88**, 1007–1019.

P. Congdon (2006). Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Computational Statistics and Data Analysis*, **50**, 346–357.

R. J. Cook, E. T. M. Ng, and M. O. Meade (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. *Biometrics*, **56**, 1109–1117.

A. C. Davison and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, MA.

C. M. Dayton and G. B. Macready (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, **83**, 173–178.

C. M. Dayton and G. B. Macready (2002). Use of categorical and continuous covariates in latent class analysis. In: J. A. Hagenaars and A. L. McCutcheon, eds., *Advances in Latent Class Modeling*, 213–233. Cambridge University Press, Cambridge, MA.

J. De Leeuw and N. Verhelst (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, **11**, 183–196.

A. P. Dempster, N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

J. G. Dias (2006). Model selection for the binary latent class model: a Monte Carlo simulation. In: A. Ferligoj, V. Batagelj, H.-H. Bock and A. Ziberna, eds., *Data Science and Classification*, 91–100. Springer, Berlin.

J. G. Dias and J. K. Vermunt (2007). Latent class modeling of website users' search patterns: implications for online market segmentation. *Journal of Retailing and Consumer Services*, **14**, 359–368.

J. Diebolt and C. P. Robert (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 363–375.

P. J. Diggle, P. J. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data, 2nd Edition*. Oxford University Press, Oxford, UK.

A. V. D'Unger, K. C. Lund, P. L. McCall, and D. S. Nagin (1998). How many latent classes of delinquent/criminal careers? Results from mixed Poisson regression analyses. *American Journal of Sociology*, **103**, 1593–1630.

J. A. du Preez (1997). *Efficient High-Order Hidden Markov Modelling*. Ph.D. thesis, University of Stellenbosch.

P. Dymarski (2011). *Hidden Markov Models, Theory and Applications*. InTech, Rijeka, Croatia.

B. Efron and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York.

D. S. Elliot, D. Huizinga, and S. Menard (1989). *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems*. Springer-Verlag, New York.

R. J. Elliott, L. Aggoun, and J. B. Moore (1995). *Hidden Markov Models: Estimation and Control*, Springer, New York.

Y. Ephraim and N. Merhav (2002). Hidden Markov processes. *IEEE Transactions on Information Theory*, **48**, 1518–1569.

A. Farcomeni (2012). Quantile regression for longitudinal data based on latent Markov subject-specific parameters. *Statistics and Computing*, **22**, 141–152.

Z. D. Feng and C. E. McCulloch (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society, Series B*, **58**, 609–617.

G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (2009). *Longitudinal Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL.

A. Forcina (2008). Identifiability of extended latent class models with individual covariates. ~~*Computational Statistics and Data Analysis*~~, **52**, 5263–5268.

A. K. Formann (1995). Linear logistic latent class analysis and the Rasch model. In: G. H. Fischer and I. W. Molenaar, eds., *Rasch Models: Foundations, Recent Developments, and Applications*, 239–255. Springer, New York.

B. Francis, K. Soothill, and R. Fligelstone (2004). Identifying patterns and pathways of offending behaviour: a new approach to typologies of crime. *European Journal of Criminology*, **1**, 47–87.

E. W. Frees (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge University Press, Cambridge, UK.

S. Frühwirth-Schnatter (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**, 194–209.

M. Geraci and M. Bottai (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, **8**, 140–154.

Z. Ghahramani and M. I. Jordan (1997). Factorial hidden Markov models. *Machine Learning*, **29**, 245–273.

J. K. Ghosh, M. Delampady, and T. Samanta (2010). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer-Verlag, New York.

G. F. V. Glonek (1996). A class of regression models for multivariate categorical responses. *Biometrika*, **83**, 15–28.

G. F. V. Glonek and P. McCullagh (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, **57**, 533–546.

H. Goldstein, G. Bonnet, and T. Rocher (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, **32**, 252–286.

L. A. Goodman (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, **56**, 841–868.

L. A. Goodman (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

L. A. Goodman (2002). Latent class analysis. In: A. L. McCutCheon and J. A. Hagenaars, eds., *Applied Latent Class Analysis*. Cambridge University Press, Cambridge, MA.

B. F. Green (1951). A general solution for the latent class model of latent structure analysis. *Psychometrika*, **16**, 151–166.

P. J. Green (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

G. R. Grimmett and D. R. Stirzaker (2001). *Probability and Random Processes, 3rd Edition*. Oxford University Press, Oxford.

S. J. Haberman (1974). Log-linear models for frequency tables derived by indirected observation: maximum-likelihood equations. *Annals of Statistics*, **2**, 911–924.

R. K. Hambleton and H. Swaminathan (1985). *Item Response Theory: Principles and Applications*. Kluwer Nijhoff, Boston.

J. D. Hamilton (1989). A new approach to the economic-analysis of nonstationary time-series and the business cycle. *Econometrica*, **57**, 357–384.

W. K. Hastings (1970). Monte Carlo sampling methods using Markov chais and their applications. *Biometrika*, **57**, 97–109.

J. J. Heckman (1981). Heterogeneity and state dependence. In: D. L. McFadden and C. A. Manski, eds., *Structural Analysis of Discrete Data*. MIT Press, Cambridge, MA.

F. Heiss (2008). Sequential numerical integration in nonlinear state space models for microeconometric panel data. *Journal of Applied Econometrics*, **23**, 373–389.

P. W. Holland and P. R. Rosenbaum (1986). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, **59**, 147–177.

C. Hsiao (2003). *Analysis of Panel Data, 2nd Edition*. Cambridge University Press, Cambridge, MA.

G.-H. Huang and K. Bandeen-Roche (2004). Building an identifiable latent class model, with covariate effects on underlying and measured variables. *Psychometrika*, **69**, 5–32.

K. Humphreys and H. Janson (2000). Latent transition analysis with covariates, nonresponse, summary statistics and diagnostics: modelling children's drawing development. *Multivariate Behavioral Research*, **35**, 89–118.

M. Hurn, A. Justel, and C. P. Robert (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, **12**, 55–79.

D. R. Hyslop (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica*, **67**, 1255–1294.

A. Jasra, C. C. Holmes, and D. A. Stephens (2005). Markov chain Monte Carlo and the label switching problem in Bayesian mixture models. *Statistical Science*, **20**, 50–67.

B. H. Juang and L. R. Rabiner (1991). Hidden Markov models for speech recognition. *Technometrics*, **33**, 251–272.

S.-H. Jung (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, **91**, 251–257.

D. Kaplan (2008). An overview of Markov chain methods for the study of stage-sequential developmental processes. *Developmental Psychology*, **44**, 457–467.

R. E. Kass and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

R. Koenker (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91**, 74–89.

R. Koenker (2005). *Quantile Regression*. Cambridge University Press, New York.

R. Koenker and G. Bassett Jr. (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

T. Koski (2001). *Hidden Markov Models for Bioinformatics*. Kluwer, Dordrecht, NL.

G. L. Kouemou (2011). History and theoretical basics of hidden Markov models. In: P. Dymarski, ed., *Hidden Markov Models, Theory and Applications*, 3–26. InTech, Rijeka, Croatia.

J. B. Lang, J. W. McDonald, and P. W. F. Smith (1999). Association-marginal modeling of multivariate categorical responses: a maximum likelihood approach. *Journal of the American Statistical Association*, **94**, 1161–71.

R. Langeheine (1988). New development in latent class theory. In: R. Langeheine and J. Rost, eds., *Latent Trait and Latent Class Models*, 77–108. Plenum Press, New York.

R. Langeheine (1994). Latent variables Markov models. In: A. von Eye and C. C. Clogg, eds., *Latent Variables Analysis: Applications for Developmental Research*, 373–395. Sage, Thousand Oaks, CA.

R. Langeheine and F. Van de Pol (1994). Discrete-time mixed Markov latent class models. In: A. Dale and R. B. Davies, eds., *Analyzing Social and Political Change: A Casebook of Methods*, 171–197. Sage Publications, London.

P. F. Lazarsfeld (1950). The logical and mathematical foundation of latent structure analysis. In: E. A. Suchman, S. A. Stouffer, and L. Guttman, eds., *Measurement and Prediction*. Princeton University Press, New York.

P. F. Lazarsfeld and N. W. Henry (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

T. Leonard (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B*, **37**, 23–37.

S. E. Levinson, L. R. Rabiner, and M. M. Sondhi (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell System Technical Journal*, **62**, 1035–1074.

B. Lindsay, C. Clogg, and J. Grego (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, **86**, 96–107.

B. G. Lindsay (1995). *Mixture Models: Theory, Geometry and Applications*. Institute of Mathematical Statistics, Hayward, CA.

S. R. Lipsitz, G. M. Fitzmaurice, G. Molenberghs, and L. P. Zhao (1997). Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus. *Journal of the Royal Statistical Society, Series C*, **46**, 463–476.

R. J. A. Little and D. B. Rubin (2002). *Statistical Analysis with Missing Data, 2nd Edition*. Wiley, New York.

Y. Liu and M. Bottai (2009). Mixed-effects models for conditional quantiles with longitudinal data. *The International Journal of Biostatistics*, **5**.

T. A. Louis (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226–233.

T. C. Lystig and J. P. Hughes (2002). Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics*, **11**, 678–689.

I. L. MacDonald and W. Zucchini (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall, London.

J. Magidson and J. K. Vermunt (2001). Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology*, **31**, 223–264.

J. M. Marin, K. L. Mengersen, and C. P. Robert (2005). Bayesian modelling and inference on mixture of distributions. In: D. Dey and C. R. Rao, eds., *Handbooks of Statistics*, vol. 25, 459–507. Elsevier Science, Amsterdam.

A. Maruotti (2011). Mixed hidden Markov models for longitudinal data: an overview. *International Statistical Review*, **79**, 427–454.

G. N. Masters (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149–174.

P. McCullagh (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 109–142.

P. McCullagh and J. A. Nelder (1989). *Generalized Linear Models, 2nd Edition*. Chapman and Hall/CRC, London.

C. E. McCulloch, S. R. Searle, and J. M. Neuhaus (2008). *Generalized, Linear, and Mixed Models*. Wiley, Hoboken, NJ.

R. B. McHugh (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika*, **21**, 331–347.

G. J. McLachlan and D. Peel (2000). *Finite Mixture Models*. Wiley, New York.

G. J. McLachlan and T. Krishnan (2008). *The EM Algorithm and Extensions: 2nd Edition*. Wiley, Hoboken, NJ.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machine. *Chemical Physics*, **21**, 1087–1092.

T. Moffitt (1993). Adolescent-limited and life-course-persistent antisocial behavior: a developmental taxonomy. *Psychological Review*, **100**, 674–701.

B. Muthén (2004). Latent variable analysis: growth mixture modeling and related techniques for longitudinal data. In: D. Kaplan, ed., *Handbook of Quantitative Methodology for the Social sciences*, 345–368. Sage Publications, Newbury Park, CA.

B. Muthén and K. Shedden (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55**, 463–469.

D. Nagin (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, **4**, 139–157.

D. S. Nagin and K. C. Land (1993). Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric, mixed-Poisson model. *Criminology*, **31**, 327–362.

W. A. Nazaret (1987). Bayesian log linear estimates for three-way contingency tables. *Biometrika*, **74**, 401–410.

D. Oakes (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **61**, 479–482.

L. J. Paas, J. K. Vermunt, and T. H. A. Bilmolt (2007). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society, Series A*, **170**, 955–974.

C. S. Poulsen (1990). Mixed Markov and latent Markov modeling applied to brand choice data. *International Journal of Research in Marketing*, **7**, 5–19.

J. Prime, S. White, S. Liriano, and K. Patel (2001). Criminal careers of those born between 1953 and 1978. vol. 10. Home Office, London.

C. R. Rao (1973). *Linear Statistical Inference and Its Applications, 2nd Edition.* Wiley, New York.

G. Rasch (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 321–333.

Research Development and Statistics Directorate (1998). *The Offenders Index: Codebook.* Home Office, London.

K. Richardson, D. Harte, and K. Carter (2011). Understanding health and labour force transitions: applying Markov models to SoFIE longitudinal data. *Tech. Rep. 2011–2*, Official Statistics Research Series.

S. Richardson and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.

F. Rijmen, K. Vansteelandt, and P. De Boeck (2008). Latent class models for diary methods data: parameter estimation by local computations. *Psychometrika*, **73**, 167–182.

C. P. Robert and G. Casella (2010). *Monte Carlo Statistical Methods, 2nd Edition*. Springer-Verlag, New York.

C. P. Robert, T. Rydén, and D. M. Titterington (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B*, **62**, 57–75.

T. J. Rothenberg (1971). Identification in parametric models. *Econometrica*, **39**, 577–591.

F. Samejima (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph n. **17**, Psychometric Society, Richmond, VA.

G. Schwarz (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

S. L. Scott (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337–351.

S. G. Self and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

A. Shapiro (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review*, **56**, 49–62.

J. Q. Shi, R. Murray-Smith, and D. M. Titterington (2002). Birth-death MCMC methods for mixtures with an unknown number of components. *Tech. Rep. TR-2002-117*, Department of Computing Science, University of Glasgow.

N.-Z. Shi, S.-R. Zheng, and J. Guo (2005). The restricted EM algorithm under inequality restrictions on the parameters. *Journal of Multivariate Analysis*, **92**, 53–76.

M. J. Silvapulle and P. K. Sen (2004). *Constrained Statistical Inference*. Wiley, New York.

A. Skrondal and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modelling: Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall, Boca Raton, FL.

K. Soothill, B. Francis, and R. Fligelstone (2002). Patterns of offending behaviour: a new approach. http://www.homeoffice.gov.uk/rds/pdfs2/patternsrevisedr171.pdf

L. Spezia (2009). Reversible jump and the label switching problem in hidden Markov models. *Journal of Statistical Planning and Inference*, **139**, 2305–2315.

L. Spezia (2010). Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis*, **31**, 1–11.

D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, Chichester.

M. Stephens (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, **62**, 795–809.

H. M. Taylor and S. Karlin (1998). *An Introduction to Stochastic Modelling, 3rd Edition*. Academic Press, San Diego, CA.

D. M. Titterington, A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

R. Turner (2008). Direct maximization of the likelihood of a hidden Markov model. *Computational Statistics and Data Analysis*, **52**, 4147–4160.

F. Tuyl, R. Gerlach, and K. Mengersen (2009). Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. *Bayesian Analysis*, **4**, 151–158.

F. van de Pol and R. Langeheine (1990). Mixed Markov latent class models. *Sociological Methodology*, **20**, 213–247.

T. van Erven, P. Grünwald, and S. de Roij (2012). Catching up faster by switching sooner: a predictive approach to adoptive estimation with an application to the AIC-BIC dilemma. *Journal of the Royal Statistical Society, Series B*, **74**, 361–417.

J. K. Vermunt (2010). Longitudinal research with latent variables. In: K. van Montfort, J. H. L. Oud, and A. Satorra, eds., *Handbook of Advanced Multilevel Analysis*, 119–152. Springer, Heidelberg, Germany.

J. K. Vermunt and J. A. Hagenaars (2004). Ordinal longitudinal data analysis. In: N. Cameron, R. C. Hauspie, and L. Molinari, eds., *Methods in Human Growth Research*, pp. 374–393. Cambridge University Press.

J. K. Vermunt, R. Langeheine, and U. Böckenholt (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, **24**, 179–207.

A. J. Viterbi (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**, 260–269.

A. Wald (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistic*, **20**, 595–601.

M. Watanabe and K. Yamaguchi (2004). *The EM Algorithm and Related Statistical Models*. Marcel Dekker, New York.

L. R. Welch (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, **53**, 1–13.

L. M. Wiggins (1955). *Mathematical Models for the Analysis of Multiwave Panels.* In: Ph.D. Dissertation, Columbia University, Ann Arbor, MI.

L. M. Wiggins (1973). *Panel Analysis: Latent Probability Models for Attitude and Behaviour Processes*. Elsevier, Amsterdam.

Y. Yang (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, **92**, 937–950.

H.-T. Yu (2008). *Multilevel Latent Markov Models for Nested Longitudinal Discrete Data*. Ph.D. Dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.

K. Yu, Z. Lu, and J. Stander (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society, Series D*, **52**, 331–350.

W. Zucchini and P. Guttorp (1991). A hidden Markov model for space-time precipitation. *Water Resources Research*, **27**, 1917–1923.

W. Zucchini and I. L. MacDonald (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC, Boca Raton, FL.

Index